

1

Natural Language Processing

2

Jacob Eisenstein

3

October 12, 2018

⁴ Contents

⁵	Contents	1
⁶	Preface	i
⁷	Background	i
⁸	How to use this book	ii
⁹	1 Introduction	1
¹⁰	1.1 Natural language processing and its neighbors	1
¹¹	1.2 Three themes in natural language processing	6
¹²	1.2.1 Learning and knowledge	6
¹³	1.2.2 Search and learning	7
¹⁴	1.2.3 Relational, compositional, and distributional perspectives	9
¹⁵	I Learning	11
¹⁶	2 Linear text classification	13
¹⁷	2.1 The bag of words	13
¹⁸	2.2 Naïve Bayes	17
¹⁹	2.2.1 Types and tokens	19
²⁰	2.2.2 Prediction	20
²¹	2.2.3 Estimation	21
²²	2.2.4 Smoothing	22
²³	2.2.5 Setting hyperparameters	23
²⁴	2.3 Discriminative learning	24
²⁵	2.3.1 Perceptron	25
²⁶	2.3.2 Averaged perceptron	27
²⁷	2.4 Loss functions and large-margin classification	27
²⁸	2.4.1 Online large margin classification	30
²⁹	2.4.2 *Derivation of the online support vector machine	32
³⁰	2.5 Logistic regression	35

31	2.5.1	Regularization	36
32	2.5.2	Gradients	37
33	2.6	Optimization	37
34	2.6.1	Batch optimization	38
35	2.6.2	Online optimization	39
36	2.7	*Additional topics in classification	41
37	2.7.1	Feature selection by regularization	41
38	2.7.2	Other views of logistic regression	41
39	2.8	Summary of learning algorithms	43
40	3	Nonlinear classification	47
41	3.1	Feedforward neural networks	48
42	3.2	Designing neural networks	50
43	3.2.1	Activation functions	50
44	3.2.2	Network structure	51
45	3.2.3	Outputs and loss functions	52
46	3.2.4	Inputs and lookup layers	53
47	3.3	Learning neural networks	53
48	3.3.1	Backpropagation	55
49	3.3.2	Regularization and dropout	57
50	3.3.3	*Learning theory	58
51	3.3.4	Tricks	59
52	3.4	Convolutional neural networks	62
53	4	Linguistic applications of classification	69
54	4.1	Sentiment and opinion analysis	69
55	4.1.1	Related problems	70
56	4.1.2	Alternative approaches to sentiment analysis	72
57	4.2	Word sense disambiguation	73
58	4.2.1	How many word senses?	74
59	4.2.2	Word sense disambiguation as classification	75
60	4.3	Design decisions for text classification	76
61	4.3.1	What is a word?	76
62	4.3.2	How many words?	79
63	4.3.3	Count or binary?	80
64	4.4	Evaluating classifiers	80
65	4.4.1	Precision, recall, and <i>F</i> -MEASURE	81
66	4.4.2	Threshold-free metrics	83
67	4.4.3	Classifier comparison and statistical significance	84
68	4.4.4	*Multiple comparisons	87
69	4.5	Building datasets	88

70	4.5.1 Metadata as labels	88
71	4.5.2 Labeling data	88
72	5 Learning without supervision	95
73	5.1 Unsupervised learning	95
74	5.1.1 K -means clustering	96
75	5.1.2 Expectation-Maximization (EM)	98
76	5.1.3 EM as an optimization algorithm	102
77	5.1.4 How many clusters?	103
78	5.2 Applications of expectation-maximization	104
79	5.2.1 Word sense induction	104
80	5.2.2 Semi-supervised learning	105
81	5.2.3 Multi-component modeling	106
82	5.3 Semi-supervised learning	107
83	5.3.1 Multi-view learning	108
84	5.3.2 Graph-based algorithms	109
85	5.4 Domain adaptation	110
86	5.4.1 Supervised domain adaptation	111
87	5.4.2 Unsupervised domain adaptation	112
88	5.5 *Other approaches to learning with latent variables	114
89	5.5.1 Sampling	115
90	5.5.2 Spectral learning	117
91	II Sequences and trees	123
92	6 Language models	125
93	6.1 N -gram language models	126
94	6.2 Smoothing and discounting	129
95	6.2.1 Smoothing	129
96	6.2.2 Discounting and backoff	130
97	6.2.3 *Interpolation	131
98	6.2.4 *Kneser-Ney smoothing	133
99	6.3 Recurrent neural network language models	133
100	6.3.1 Backpropagation through time	136
101	6.3.2 Hyperparameters	137
102	6.3.3 Gated recurrent neural networks	137
103	6.4 Evaluating language models	139
104	6.4.1 Held-out likelihood	139
105	6.4.2 Perplexity	140
106	6.5 Out-of-vocabulary words	141

107	7 Sequence labeling	145
108	7.1 Sequence labeling as classification	145
109	7.2 Sequence labeling as structure prediction	147
110	7.3 The Viterbi algorithm	149
111	7.3.1 Example	152
112	7.3.2 Higher-order features	153
113	7.4 Hidden Markov Models	153
114	7.4.1 Estimation	155
115	7.4.2 Inference	155
116	7.5 Discriminative sequence labeling with features	157
117	7.5.1 Structured perceptron	160
118	7.5.2 Structured support vector machines	160
119	7.5.3 Conditional random fields	162
120	7.6 Neural sequence labeling	167
121	7.6.1 Recurrent neural networks	167
122	7.6.2 Character-level models	169
123	7.6.3 Convolutional Neural Networks for Sequence Labeling	170
124	7.7 *Unsupervised sequence labeling	170
125	7.7.1 Linear dynamical systems	172
126	7.7.2 Alternative unsupervised learning methods	172
127	7.7.3 Semiring notation and the generalized viterbi algorithm	172
128	8 Applications of sequence labeling	175
129	8.1 Part-of-speech tagging	175
130	8.1.1 Parts-of-Speech	176
131	8.1.2 Accurate part-of-speech tagging	180
132	8.2 Morphosyntactic Attributes	182
133	8.3 Named Entity Recognition	183
134	8.4 Tokenization	185
135	8.5 Code switching	186
136	8.6 Dialogue acts	187
137	9 Formal language theory	191
138	9.1 Regular languages	192
139	9.1.1 Finite state acceptors	193
140	9.1.2 Morphology as a regular language	194
141	9.1.3 Weighted finite state acceptors	196
142	9.1.4 Finite state transducers	201
143	9.1.5 *Learning weighted finite state automata	206
144	9.2 Context-free languages	207
145	9.2.1 Context-free grammars	208

146	9.2.2 Natural language syntax as a context-free language	211
147	9.2.3 A phrase-structure grammar for English	213
148	9.2.4 Grammatical ambiguity	218
149	9.3 *Mildly context-sensitive languages	218
150	9.3.1 Context-sensitive phenomena in natural language	219
151	9.3.2 Combinatory categorial grammar	220
152	10 Context-free parsing	225
153	10.1 Deterministic bottom-up parsing	226
154	10.1.1 Recovering the parse tree	227
155	10.1.2 Non-binary productions	227
156	10.1.3 Complexity	229
157	10.2 Ambiguity	229
158	10.2.1 Parser evaluation	230
159	10.2.2 Local solutions	231
160	10.3 Weighted Context-Free Grammars	232
161	10.3.1 Parsing with weighted context-free grammars	234
162	10.3.2 Probabilistic context-free grammars	235
163	10.3.3 *Semiring weighted context-free grammars	237
164	10.4 Learning weighted context-free grammars	238
165	10.4.1 Probabilistic context-free grammars	238
166	10.4.2 Feature-based parsing	239
167	10.4.3 *Conditional random field parsing	240
168	10.4.4 Neural context-free grammars	242
169	10.5 Grammar refinement	242
170	10.5.1 Parent annotations and other tree transformations	243
171	10.5.2 Lexicalized context-free grammars	244
172	10.5.3 *Refinement grammars	248
173	10.6 Beyond context-free parsing	249
174	10.6.1 Reranking	250
175	10.6.2 Transition-based parsing	251
176	11 Dependency parsing	257
177	11.1 Dependency grammar	257
178	11.1.1 Heads and dependents	258
179	11.1.2 Labeled dependencies	259
180	11.1.3 Dependency subtrees and constituents	260
181	11.2 Graph-based dependency parsing	262
182	11.2.1 Graph-based parsing algorithms	264
183	11.2.2 Computing scores for dependency arcs	265
184	11.2.3 Learning	267

185	11.3 Transition-based dependency parsing	268
186	11.3.1 Transition systems for dependency parsing	269
187	11.3.2 Scoring functions for transition-based parsers	273
188	11.3.3 Learning to parse	274
189	11.4 Applications	277
190	III Meaning	283
191	12 Logical semantics	285
192	12.1 Meaning and denotation	286
193	12.2 Logical representations of meaning	287
194	12.2.1 Propositional logic	287
195	12.2.2 First-order logic	288
196	12.3 Semantic parsing and the lambda calculus	291
197	12.3.1 The lambda calculus	292
198	12.3.2 Quantification	293
199	12.4 Learning semantic parsers	296
200	12.4.1 Learning from derivations	297
201	12.4.2 Learning from logical forms	299
202	12.4.3 Learning from denotations	301
203	13 Predicate-argument semantics	305
204	13.1 Semantic roles	307
205	13.1.1 VerbNet	308
206	13.1.2 Proto-roles and PropBank	309
207	13.1.3 FrameNet	310
208	13.2 Semantic role labeling	312
209	13.2.1 Semantic role labeling as classification	312
210	13.2.2 Semantic role labeling as constrained optimization	315
211	13.2.3 Neural semantic role labeling	317
212	13.3 Abstract Meaning Representation	318
213	13.3.1 AMR Parsing	321
214	14 Distributional and distributed semantics	325
215	14.1 The distributional hypothesis	325
216	14.2 Design decisions for word representations	327
217	14.2.1 Representation	327
218	14.2.2 Context	328
219	14.2.3 Estimation	329
220	14.3 Latent semantic analysis	329

221	14.4 Brown clusters	331
222	14.5 Neural word embeddings	334
223	14.5.1 Continuous bag-of-words (CBOW)	334
224	14.5.2 Skipgrams	335
225	14.5.3 Computational complexity	335
226	14.5.4 Word embeddings as matrix factorization	337
227	14.6 Evaluating word embeddings	338
228	14.6.1 Intrinsic evaluations	339
229	14.6.2 Extrinsic evaluations	339
230	14.6.3 Fairness and bias	340
231	14.7 Distributed representations beyond distributional statistics	341
232	14.7.1 Word-internal structure	341
233	14.7.2 Lexical semantic resources	343
234	14.8 Distributed representations of multiword units	344
235	14.8.1 Purely distributional methods	344
236	14.8.2 Distributional-compositional hybrids	345
237	14.8.3 Supervised compositional methods	346
238	14.8.4 Hybrid distributed-symbolic representations	346
239	15 Reference Resolution	351
240	15.1 Forms of referring expressions	352
241	15.1.1 Pronouns	352
242	15.1.2 Proper Nouns	357
243	15.1.3 Nominals	357
244	15.2 Algorithms for coreference resolution	358
245	15.2.1 Mention-pair models	359
246	15.2.2 Mention-ranking models	360
247	15.2.3 Transitive closure in mention-based models	361
248	15.2.4 Entity-based models	362
249	15.3 Representations for coreference resolution	367
250	15.3.1 Features	367
251	15.3.2 Distributed representations of mentions and entities	370
252	15.4 Evaluating coreference resolution	373
253	16 Discourse	379
254	16.1 Segments	379
255	16.1.1 Topic segmentation	380
256	16.1.2 Functional segmentation	381
257	16.2 Entities and reference	381
258	16.2.1 Centering theory	382
259	16.2.2 The entity grid	383

260	16.2.3 *Formal semantics beyond the sentence level	384
261	16.3 Relations	385
262	16.3.1 Shallow discourse relations	385
263	16.3.2 Hierarchical discourse relations	389
264	16.3.3 Argumentation	392
265	16.3.4 Applications of discourse relations	393
266	IV Applications	401
267	17 Information extraction	403
268	17.1 Entities	405
269	17.1.1 Entity linking by learning to rank	406
270	17.1.2 Collective entity linking	408
271	17.1.3 *Pairwise ranking loss functions	409
272	17.2 Relations	411
273	17.2.1 Pattern-based relation extraction	412
274	17.2.2 Relation extraction as a classification task	413
275	17.2.3 Knowledge base population	416
276	17.2.4 Open information extraction	419
277	17.3 Events	420
278	17.4 Hedges, denials, and hypotheticals	422
279	17.5 Question answering and machine reading	424
280	17.5.1 Formal semantics	424
281	17.5.2 Machine reading	425
282	18 Machine translation	431
283	18.1 Machine translation as a task	431
284	18.1.1 Evaluating translations	433
285	18.1.2 Data	435
286	18.2 Statistical machine translation	436
287	18.2.1 Statistical translation modeling	437
288	18.2.2 Estimation	438
289	18.2.3 Phrase-based translation	439
290	18.2.4 *Syntax-based translation	441
291	18.3 Neural machine translation	442
292	18.3.1 Neural attention	444
293	18.3.2 *Neural machine translation without recurrence	446
294	18.3.3 Out-of-vocabulary words	448
295	18.4 Decoding	449
296	18.5 Training towards the evaluation metric	451

297	19 Text generation	457
298	19.1 Data-to-text generation	457
299	19.1.1 Latent data-to-text alignment	459
300	19.1.2 Neural data-to-text generation	460
301	19.2 Text-to-text generation	464
302	19.2.1 Neural abstractive summarization	464
303	19.2.2 Sentence fusion for multi-document summarization	465
304	19.3 Dialogue	466
305	19.3.1 Finite-state and agenda-based dialogue systems	467
306	19.3.2 Markov decision processes	468
307	19.3.3 Neural chatbots	470
308	A Probability	475
309	A.1 Probabilities of event combinations	475
310	A.1.1 Probabilities of disjoint events	476
311	A.1.2 Law of total probability	477
312	A.2 Conditional probability and Bayes' rule	477
313	A.3 Independence	479
314	A.4 Random variables	480
315	A.5 Expectations	481
316	A.6 Modeling and estimation	482
317	B Numerical optimization	485
318	B.1 Gradient descent	486
319	B.2 Constrained optimization	486
320	B.3 Example: Passive-aggressive online learning	487
321	Bibliography	489

³²² Preface

³²³ This text began with notes that I wrote for Georgia Tech’s undergraduate and gradu-
³²⁴ ate courses on natural language processing, CS 4650 and 7650. There are several other
³²⁵ good resources (e.g., Manning and Schütze, 1999; Jurafsky and Martin, 2009; Smith, 2011;
³²⁶ Collins, 2013), but the goal of this text is focus on a core subset of the field, unified by the
³²⁷ concepts of learning and search. A remarkable thing about natural language processing
³²⁸ is that so many problems can be solved by a compact set of methods:

³²⁹ **Search.** Viterbi, CKY, minimum spanning tree, shift-reduce, integer linear programming,
³³⁰ beam search.

³³¹ **Learning.** Maximum-likelihood estimation, logistic regression, perceptron, expectation-
³³² maximization, matrix factorization, backpropagation.

³³³ This text explains how these methods work, and how they can be applied to problems
³³⁴ that arise in the computer processing of natural language: document classification, word
³³⁵ sense disambiguation, sequence labeling (part-of-speech tagging and named entity recog-
³³⁶ nition), parsing, coreference resolution, relation extraction, discourse analysis, language
³³⁷ modeling, and machine translation.

³³⁸ Background

³³⁹ Because natural language processing draws on many different intellectual traditions, al-
³⁴⁰ most everyone who approaches it feels underprepared in one way or another. Here is a
³⁴¹ summary of what is expected, and where you can learn more:

³⁴² **Mathematics and machine learning.** The text assumes a background in multivariate cal-
³⁴³ culus and linear algebra: vectors, matrices, derivatives, and partial derivatives. You
³⁴⁴ should also be familiar with probability and statistics. A review of basic proba-
³⁴⁵ bility is found in Appendix A, and a minimal review of numerical optimization is
³⁴⁶ found in Appendix B. For linear algebra, the online course and textbook from Strang
³⁴⁷ (2016) are excellent sources of review material. Deisenroth et al. (2018) are currently

348 preparing a textbook on *Mathematics for Machine Learning*, and several chapters can
349 be found online.¹ For an introduction to probabilistic modeling and estimation, see
350 James et al. (2013); for a more advanced and comprehensive discussion of the same
351 material, the classic reference is Hastie et al. (2009).

352 **Linguistics.** This book assumes no formal training in linguistics, aside from elementary
353 concepts like nouns and verbs, which you have probably encountered in the study
354 of English grammar. Ideas from linguistics are introduced throughout the text as
355 needed, including discussions of morphology and syntax (chapter 9), semantics
356 (chapters 12 and 13), and discourse (chapter 16). Linguistic issues also arise in the
357 application-focused chapters 4, 8, and 18. A short guide to linguistics for students
358 of natural language processing is offered by Bender (2013); you are encouraged to
359 start there, and then pick up a more comprehensive introductory textbook (e.g., Ak-
360 majian et al., 2010; Fromkin et al., 2013).

361 **Computer science.** The book is targeted at computer scientists, who are assumed to have
362 taken introductory courses on the analysis of algorithms and complexity theory. In
363 particular, you should be familiar with asymptotic analysis of the time and memory
364 costs of algorithms, and with the basics of dynamic programming. The classic text
365 on algorithms is offered by Cormen et al. (2009); for an introduction to the theory of
366 computation, see Arora and Barak (2009) and Sipser (2012).

367 How to use this book

368 After the introduction, the rest of the textbook is organized into four main units:

369 **Learning.** This section builds up a set of machine learning tools that will be used through-
370 out the other sections. Because the focus is on machine learning, the text repres-
371 entations and linguistic phenomena are mostly simple: “bag-of-words” text classifi-
372 cation is treated as a model example. Chapter 4 describes some of the more linguisti-
373 cally interesting applications of word-based text analysis.

374 **Sequences and trees.** This section introduces the treatment of language as a structured
375 phenomena. It describes sequence and tree representations and the algorithms that
376 they facilitate, as well as the limitations that these representations impose. Chap-
377 ter 9 introduces finite state automata and briefly overviews a context-free account of
378 English syntax.

379 **Meaning.** This section takes a broad view of efforts to represent and compute meaning
380 from text, ranging from formal logic to neural word embeddings. It also includes

¹<https://mml-book.github.io/>

381 two topics that are closely related to semantics: resolution of ambiguous references,
 382 and analysis of multi-sentence discourse structure.

383 **Applications.** The final section offers chapter-length treatments on three of the most promi-
 384 nent applications of natural language processing: information extraction, machine
 385 translation, and text generation. Each of these applications merits a textbook length
 386 treatment of its own (Koehn, 2009; Grishman, 2012; Reiter and Dale, 2000); the chap-
 387 ters here explain some of the most well known systems using the formalisms and
 388 methods built up earlier in the book, while introducing methods such as neural at-
 389 tention.

390 Each chapter contains some advanced material, which is marked with an asterisk.
 391 This material can be safely omitted without causing misunderstandings later on. But
 392 even without these advanced sections, the text is too long for a single semester course, so
 393 instructors will have to pick and choose among the chapters.

394 Chapters 1-3 provide building blocks that will be used throughout the book, and chap-
 395 ter 4 describes some critical aspects of the practice of language technology. Language
 396 models (chapter 6), sequence labeling (chapter 7), and parsing (chapter 10 and 11) are
 397 canonical topics in natural language processing, and distributed word embeddings (chap-
 398 ter 14) have become ubiquitous. Of the applications, machine translation (chapter 18) is
 399 the best choice: it is more cohesive than information extraction, and more mature than text
 400 generation. In my experience, nearly all students benefit from the review of probability in
 401 Appendix A.

- 402 • A course focusing on machine learning should add the chapter on unsupervised
 403 learning (chapter 5). The chapters on predicate-argument semantics (chapter 13),
 404 reference resolution (chapter 15), and text generation (chapter 19) are particularly
 405 influenced by recent machine learning innovations, including deep neural networks
 406 and learning to search.
- 407 • A course with a more linguistic orientation should add the chapters on applica-
 408 tions of sequence labeling (chapter 8), formal language theory (chapter 9), semantics
 409 (chapter 12 and 13), and discourse (chapter 16).
- 410 • For a course with a more applied focus, I recommend the chapters on applications
 411 of sequence labeling (chapter 8), predicate-argument semantics (chapter 13), infor-
 412 mation extraction (chapter 17), and text generation (chapter 19).

413 Acknowledgments

414 Several colleagues, students, and friends read early drafts of chapters in their areas of
 415 expertise, including Yoav Artzi, Kevin Duh, Heng Ji, Jessy Li, Brendan O'Connor, Yuval

416 Pinter, Shawn Ling Ramirez, Nathan Schneider, Pamela Shapiro, Noah A. Smith, Sandeep
417 Soni, and Luke Zettlemoyer. I also thank the anonymous reviewers, particularly reviewer
418 4, who provided detailed line-by-line edits and suggestions. The text benefited from dis-
419 cussions with my editor Marie Lufkin Lee, as well as Kevin Murphy, Shawn Ling Ramirez,
420 and Bonnie Webber. In addition, there are many students, colleagues, friends, and family
421 who found mistakes in early drafts, or who recommended key references. These include:
422 Parminder Bhatia, Kimberly Caras, Jiahao Cai, Justin Chen, Murtaza Dhuliawala, Yantao
423 Du, Barbara Eisenstein, Luiz C. F. Ribeiro, Chris Gu, Joshua Killingsworth, Jonathan May,
424 Taha Merghani, Gus Monod, Raghavendra Murali, Nidish Nair, Brendan O'Connor, Bran-
425 don Peck, Yuval Pinter, Nathan Schneider, Jianhao Shen, Zhewei Sun, Rubin Tsui, Ashwin
426 Cunnappakkam Vinjimir, Denny Vrandečić, William Yang Wang, Clay Washington, Ishan
427 Waykul, Xavier Yao, Yuyu Zhang, and also some anonymous commenters. Clay Wash-
428 ington tested several of the programming exercises.

429 The textbook was mainly written at Georgia Tech’s School of Interactive Computing,
430 in part through teaching CS 4650 and 7650. I thank the School for its support of this
431 project, and more generally, for creating a dynamic and supportive environment for junior
432 faculty. I also thank (and apologize to) the many students in Georgia Tech’s CS 4650 and
433 7650 who suffered through early versions of the text. The book is dedicated to my parents.

⁴³⁴ Notation

As a general rule, words, word counts, and other types of observations are indicated with Roman letters (a, b, c); parameters are indicated with Greek letters (α, β, θ). Vectors are indicated with bold script for both random variables \mathbf{x} and parameters $\boldsymbol{\theta}$. Other useful notations are indicated in the table below.

Basics

$\exp x$	the base-2 exponent, 2^x
$\log x$	the base-2 logarithm, $\log_2 x$
$\{x_n\}_{n=1}^N$	the set $\{x_1, x_2, \dots, x_N\}$
x_i^j	x_i raised to the power j
$x_i^{(j)}$	indexing by both i and j

Linear algebra

$\mathbf{x}^{(i)}$	a column vector of feature counts for instance i , often word counts
$\mathbf{x}_{j:k}$	elements j through k (inclusive) of a vector \mathbf{x}
$[\mathbf{x}; \mathbf{y}]$	vertical concatenation of two column vectors
$[\mathbf{x}, \mathbf{y}]$	horizontal concatenation of two column vectors
\mathbf{e}_n	a “one-hot” vector with a value of 1 at position n , and zero everywhere else
$\boldsymbol{\theta}^\top$	the transpose of a column vector $\boldsymbol{\theta}$
$\boldsymbol{\theta} \cdot \mathbf{x}^{(i)}$	the dot product $\sum_{j=1}^N \theta_j \times x_j^{(i)}$
\mathbf{X}	a matrix
$x_{i,j}$	row i , column j of matrix \mathbf{X}
$\text{Diag}(\mathbf{x})$	a matrix with \mathbf{x} on the diagonal, e.g., $\begin{pmatrix} x_1 & 0 & 0 \\ 0 & x_2 & 0 \\ 0 & 0 & x_3 \end{pmatrix}$
\mathbf{X}^{-1}	the inverse of matrix \mathbf{X}

Text datasets

w_m	word token at position m
N	number of training instances
M	length of a sequence (of words or tags)
V	number of words in vocabulary
$y^{(i)}$	the true label for instance i
\hat{y}	a predicted label
\mathcal{Y}	the set of all possible labels
K	number of possible labels $K = \mathcal{Y} $
\square	the start token
\blacksquare	the stop token
$\mathbf{y}^{(i)}$	a structured label for instance i , such as a tag sequence
$\mathcal{Y}(\mathbf{w})$	the set of possible labelings for the word sequence \mathbf{w}
\diamond	the start tag
\blacklozenge	the stop tag

Probabilities

$\Pr(A)$	probability of event A
$\Pr(A B)$	probability of event A , conditioned on event B
$p_B(b)$	the marginal probability of random variable B taking value b ; written $p(b)$ when the choice of random variable is clear from context
$p_{B A}(b a)$	the probability of random variable B taking value b , conditioned on A taking value a ; written $p(b a)$ when clear from context
$A \sim p$	the random variable A is distributed according to distribution p . For example, $X \sim \mathcal{N}(0, 1)$ states that the random variable X is drawn from a normal distribution with zero mean and unit variance.
$A B \sim p$	conditioned on the random variable B , A is distributed according to p . ²

Machine learning

$\Psi(\mathbf{x}^{(i)}, y)$	the score for assigning label y to instance i
$\mathbf{f}(\mathbf{x}^{(i)}, y)$	the feature vector for instance i with label y
θ	a (column) vector of weights
$\ell^{(i)}$	loss on an individual instance i
L	objective function for an entire dataset
\mathcal{L}	log-likelihood of a dataset
λ	the amount of regularization

⁴³⁹ **Chapter 1**

⁴⁴⁰ **Introduction**

⁴⁴¹ Natural language processing is the set of methods for making human language accessible
⁴⁴² to computers. In the past decade, natural language processing has become embedded
⁴⁴³ in our daily lives: automatic machine translation is ubiquitous on the web and in social
⁴⁴⁴ media; text classification keeps emails from collapsing under a deluge of spam; search
⁴⁴⁵ engines have moved beyond string matching and network analysis to a high degree of
⁴⁴⁶ linguistic sophistication; dialog systems provide an increasingly common and effective
⁴⁴⁷ way to get and share information.

⁴⁴⁸ These diverse applications are based on a common set of ideas, drawing on algo-
⁴⁴⁹ rithms, linguistics, logic, statistics, and more. The goal of this text is to provide a survey
⁴⁵⁰ of these foundations. The technical fun starts in the next chapter; the rest of this current
⁴⁵¹ chapter situates natural language processing with respect to other intellectual disciplines,
⁴⁵² identifies some high-level themes in contemporary natural language processing, and ad-
⁴⁵³ vises the reader on how best to approach the subject.

⁴⁵⁴ **1.1 Natural language processing and its neighbors**

⁴⁵⁵ Natural language processing draws on many other intellectual traditions, from formal
⁴⁵⁶ linguistics to statistical physics. This section briefly situates natural language processing
⁴⁵⁷ with respect to some of its closest neighbors.

⁴⁵⁸ **Computational Linguistics** Most of the meetings and journals that host natural lan-
⁴⁵⁹ guage processing research bear the name “computational linguistics”, and the terms may
⁴⁶⁰ be thought of as essentially synonymous. But while there is substantial overlap, there is
⁴⁶¹ an important difference in focus. In linguistics, language is the object of study. Compu-
⁴⁶² tational methods may be brought to bear, just as in scientific disciplines like computational
⁴⁶³ biology and computational astronomy, but they play only a supporting role. In contrast,

464 natural language processing is focused on the design and analysis of computational al-
 465 gorithms and representations for processing natural human language. The goal of natu-
 466 ral language processing is to provide new computational capabilities around human lan-
 467 guage: for example, extracting information from texts, translating between languages, an-
 468 swering questions, holding a conversation, taking instructions, and so on. Fundamental
 469 linguistic insights may be crucial for accomplishing these tasks, but success is ultimately
 470 measured by whether and how well the job gets done.

471 **Machine Learning** Contemporary approaches to natural language processing rely heav-
 472 ily on machine learning, which makes it possible to build complex computer programs
 473 from examples. Machine learning provides an array of general techniques for tasks like
 474 converting a sequence of discrete tokens in one vocabulary to a sequence of discrete to-
 475 kens in another vocabulary — a generalization of what one might informally call “transla-
 476 tion.” Much of today’s natural language processing research can be thought of as applied
 477 machine learning. However, natural language processing has characteristics that distin-
 478 guish it from many of machine learning’s other application domains.

- 479 • Unlike images or audio, text data is fundamentally discrete, with meaning created
 480 by combinatorial arrangements of symbolic units. This is particularly consequential
 481 for applications in which text is the output, such as translation and summarization,
 482 because it is not possible to gradually approach an optimal solution.
 - 483 • Although the set of words is discrete, new words are always being created. Further-
 484 more, the distribution over words (and other linguistic elements) resembles that of a
 485 **power law**¹ (Zipf, 1949): there will be a few words that are very frequent, and a long
 486 tail of words that are rare. A consequence is that natural language processing algo-
 487 rithms must be especially robust to observations that do not occur in the training
 488 data.
 - 489 • Language is **compositional**: units such as words can combine to create phrases,
 490 which can combine by the very same principles to create larger phrases. For ex-
 491 ample, a **noun phrase** can be created by combining a smaller noun phrase with a
 492 **prepositional phrase**, as in *the whiteness of the whale*. The prepositional phrase is
 493 created by combining a preposition (in this case, *of*) with another noun phrase (*the
 494 whale*). In this way, it is possible to create arbitrarily long phrases, such as,
- 495 (1.1) ...huge globular pieces of the whale of the bigness of a human head.²

496 The meaning of such a phrase must be analyzed in accord with the underlying hier-
 497 archical structure. In this case, *huge globular pieces of the whale* acts as a single noun

¹Throughout the text, **boldface** will be used to indicate keywords that appear in the index.

²Throughout the text, this notation will be used to introduce linguistic examples.

498 phrase, which is conjoined with the prepositional phrase *of the bigness of a human*
499 *head*. The interpretation would be different if instead, *huge globular pieces* were con-
500 joined with the prepositional phrase *of the whale of the bigness of a human head* —
501 implying a disappointingly small whale. Even though text appears as a sequence,
502 machine learning methods must account for its implicit recursive structure.

503 **Artificial Intelligence** The goal of artificial intelligence is to build software and robots
504 with the same range of abilities as humans (Russell and Norvig, 2009). Natural language
505 processing is relevant to this goal in several ways. On the most basic level, the capacity for
506 language is one of the central features of human intelligence, and is therefore a prerequi-
507 site for artificial intelligence.³ Second, much of artificial intelligence research is dedicated
508 to the development of systems that can reason from premises to a conclusion, but such
509 algorithms are only as good as what they know (Dreyfus, 1992). Natural language pro-
510 cessing is a potential solution to the “knowledge bottleneck”, by acquiring knowledge
511 from texts, and perhaps also from conversations. This idea goes all the way back to Tur-
512 ing’s 1949 paper *Computing Machinery and Intelligence*, which proposed the **Turing test** for
513 determining whether artificial intelligence had been achieved (Turing, 2009).

514 Conversely, reasoning is sometimes essential for basic tasks of language processing,
515 such as resolving a pronoun. **Winograd schemas** are examples in which a single word
516 changes the likely referent of a pronoun, in a way that seems to require knowledge and
517 reasoning to decode (Levesque et al., 2011). For example,

518 (1.2) The trophy doesn’t fit into the brown suitcase because **it** is too [small/large].

519 When the final word is *small*, then the pronoun *it* refers to the suitcase; when the final
520 word is *large*, then *it* refers to the trophy. Solving this example requires spatial reasoning;
521 other schemas require reasoning about actions and their effects, emotions and intentions,
522 and social conventions.

523 Such examples demonstrate that natural language understanding cannot be achieved
524 in isolation from knowledge and reasoning. Yet the history of artificial intelligence has
525 been one of increasing specialization: with the growing volume of research in subdisci-
526 plines such as natural language processing, machine learning, and computer vision, it is

³This view is shared by some, but not all, prominent researchers in artificial intelligence. Michael Jordan, a specialist in machine learning, has said that if he had a billion dollars to spend on any large research project, he would spend it on natural language processing (https://www.reddit.com/r/MachineLearning/comments/2fxi6v/ama_michael_i_jordan/). On the other hand, in a public discussion about the future of artificial intelligence in February 2018, computer vision researcher Yann Lecun argued that despite its many practical applications, language is perhaps “number 300” in the priority list for artificial intelligence research, and that it would be a great achievement if AI could attain the capabilities of an orangutan, which do not include language (<http://www.abigailsee.com/2018/02/21/deep-learning-structure-and-innate-priors.html>).

527 difficult for anyone to maintain expertise across the entire field. Still, recent work has
528 demonstrated interesting connections between natural language processing and other ar-
529 eas of AI, including computer vision (e.g., Antol et al., 2015) and game playing (e.g.,
530 Branavan et al., 2009). The dominance of machine learning throughout artificial intel-
531 ligence has led to a broad consensus on representations such as graphical models and
532 computation graphs, and on algorithms such as backpropagation and combinatorial opti-
533 mization. Many of the algorithms and representations covered in this text are part of this
534 consensus.

535 **Computer Science** The discrete and recursive nature of natural language invites the ap-
536 plication of theoretical ideas from computer science. Linguists such as Chomsky and
537 Montague have shown how formal language theory can help to explain the syntax and
538 semantics of natural language. Theoretical models such as finite-state and pushdown au-
539 tomata are the basis for many practical natural language processing systems. Algorithms
540 for searching the combinatorial space of analyses of natural language utterances can be
541 analyzed in terms of their computational complexity, and theoretically motivated approx-
542 imations can sometimes be applied.

543 The study of computer systems is also relevant to natural language processing. Large
544 datasets of unlabeled text can be processed more quickly by parallelization techniques
545 like MapReduce (Dean and Ghemawat, 2008; Lin and Dyer, 2010); high-volume data
546 sources such as social media can be summarized efficiently by approximate streaming
547 and sketching techniques (Goyal et al., 2009). When deep neural networks are imple-
548 mented in production systems, it is possible to eke out speed gains using techniques such
549 as reduced-precision arithmetic (Wu et al., 2016). Many classical natural language process-
550 ing algorithms are not naturally suited to graphics processing unit (GPU) parallelization,
551 suggesting directions for further research at the intersection of natural language process-
552 ing and computing hardware (Yi et al., 2011).

553 **Speech Processing** Natural language is often communicated in spoken form, and speech
554 recognition is the task of converting an audio signal to text. From one perspective, this is
555 a signal processing problem, which might be viewed as a preprocessing step before nat-
556 ural language processing can be applied. However, context plays a critical role in speech
557 recognition by human listeners: knowledge of the surrounding words influences percep-
558 tion and helps to correct for noise (Miller et al., 1951). For this reason, speech recognition
559 is often integrated with text analysis, particularly with statistical **language models**, which
560 quantify the probability of a sequence of text (see chapter 6). Beyond speech recognition,
561 the broader field of speech processing includes the study of speech-based dialogue sys-
562 tems, which are briefly discussed in chapter 19. Historically, speech processing has often
563 been pursued in electrical engineering departments, while natural language processing

564 has been the purview of computer scientists. For this reason, the extent of interaction
565 between these two disciplines is less than it might otherwise be.

566 **Ethics** As machine learning and artificial intelligence become increasingly ubiquitous, it
567 is crucial to understand how their benefits, costs, and risks are distributed across differ-
568 ent kinds of people. Natural language processing raises some particularly salient issues
569 around **ethics, fairness, and accountability**:

570 **Access.** Who is natural language processing designed to serve? For example, whose lan-
571 guage is translated *from*, and whose language is translated *to*?

572 **Bias.** Does language technology learn to replicate social biases from text corpora, and
573 does it reinforce these biases as seemingly objective computational conclusions?

574 **Labor.** Whose text and speech comprise the datasets that power natural language pro-
575 cessing, and who performs the annotations? Are the benefits of this technology
576 shared with all the people whose work makes it possible?

577 **Privacy and internet freedom.** What is the impact of large-scale text processing on the
578 right to free and private communication? What is the potential role of natural lan-
579 guage processing in regimes of censorship or surveillance?

580 This text lightly touches on issues related to fairness and bias in § 14.6.3 and § 18.1.1,
581 but these issues are worthy of a book of their own. For more from within the field of
582 computational linguistics, see the papers from the annual workshop on Ethics in Natural
583 Language Processing (Hovy et al., 2017; Alfano et al., 2018). For an outside perspective on
584 ethical issues relating to data science at large, see boyd and Crawford (2012).

585 **Others** Natural language processing plays a significant role in emerging interdisciplinary
586 fields like **computational social science** and the **digital humanities**. Text classification
587 (chapter 4), clustering (chapter 5), and information extraction (chapter 17) are particularly
588 useful tools; another is **probabilistic topic models** (Blei, 2012), which are not covered in
589 this text. **Information retrieval** (Manning et al., 2008) makes use of similar tools, and
590 conversely, techniques such as latent semantic analysis (§ 14.3) have roots in information
591 retrieval. **Text mining** is sometimes used to refer to the application of data mining tech-
592 niques, especially classification and clustering, to text. While there is no clear distinction
593 between text mining and natural language processing (nor between data mining and ma-
594 chine learning), text mining is typically less concerned with linguistic structure, and more
595 interested in fast, scalable algorithms.

596 1.2 Three themes in natural language processing

597 Natural language processing covers a diverse range of tasks, methods, and linguistic phe-
 598 nomena. But despite the apparent incommensurability between, say, the summarization
 599 of scientific articles (§ 16.3.4) and the identification of suffix patterns in Spanish verbs
 600 (§ 9.1.4), some general themes emerge. The remainder of the introduction focuses on these
 601 themes, which will recur in various forms through the text. Each theme can be expressed
 602 as an opposition between two extreme viewpoints on how to process natural language.
 603 The methods discussed in the text can usually be placed somewhere on the continuum
 604 between these two extremes.

605 1.2.1 Learning and knowledge

606 A recurring topic of debate is the relative importance of machine learning and linguistic
 607 knowledge. On one extreme, advocates of “natural language processing from scratch” (Col-
 608 lobert et al., 2011) propose to use machine learning to train end-to-end systems that trans-
 609 mute raw text into any desired output structure: e.g., a summary, database, or transla-
 610 tion. On the other extreme, the core work of natural language processing is sometimes
 611 taken to be transforming text into a stack of general-purpose linguistic structures: from
 612 subword units called **morphemes**, to word-level **parts-of-speech**, to tree-structured repre-
 613 sentations of grammar, and beyond, to logic-based representations of meaning. In theory,
 614 these general-purpose structures should then be able to support any desired application.

615 The end-to-end approach has been buoyed by recent results in computer vision and
 616 speech recognition, in which advances in machine learning have swept away expert-
 617 engineered representations based on the fundamentals of optics and phonology (Krizhevsky
 618 et al., 2012; Graves and Jaitly, 2014). But while machine learning is an element of nearly
 619 every contemporary approach to natural language processing, linguistic representations
 620 such as syntax trees have not yet gone the way of the visual edge detector or the auditory
 621 triphone. Linguists have argued for the existence of a “language faculty” in all human be-
 622 ings, which encodes a set of abstractions specially designed to facilitate the understanding
 623 and production of language. The argument for the existence of such a language faculty
 624 is based on the observation that children learn language faster and from fewer examples
 625 than would be possible if language was learned from experience alone.⁴ From a practi-
 626 cal standpoint, linguistic structure seems to be particularly important in scenarios where
 627 training data is limited.

628 There are a number of ways in which knowledge and learning can be combined in
 629 natural language processing. Many supervised learning systems make use of carefully
 630 engineered **features**, which transform the data into a representation that can facilitate

⁴The *Language Instinct* (Pinker, 2003) articulates these arguments in an engaging and popular style. For arguments against the innateness of language, see Elman et al. (1998).

learning. For example, in a task like search, it may be useful to identify each word's **stem**, so that a system can more easily generalize across related terms such as *whale*, *whales*, *whalers*, and *whaling*. (This issue is relatively benign in English, as compared to the many other languages which include much more elaborate systems of prefixed and suffixes.) Such features could be obtained from a hand-crafted resource, like a dictionary that maps each word to a single root form. Alternatively, features can be obtained from the output of a general-purpose language processing system, such as a parser or part-of-speech tagger, which may itself be built on supervised machine learning.

Another synthesis of learning and knowledge is in model structure: building machine learning models whose architectures are inspired by linguistic theories. For example, the organization of sentences is often described as **compositional**, with meaning of larger units gradually constructed from the meaning of their smaller constituents. This idea can be built into the architecture of a deep neural network, which is then trained using contemporary deep learning techniques (Dyer et al., 2016).

The debate about the relative importance of machine learning and linguistic knowledge sometimes becomes heated. No machine learning specialist likes to be told that their engineering methodology is unscientific alchemy;⁵ nor does a linguist want to hear that the search for general linguistic principles and structures has been made irrelevant by big data. Yet there is clearly room for both types of research: we need to know how far we can go with end-to-end learning alone, while at the same time, we continue the search for linguistic representations that generalize across applications, scenarios, and languages. For more on the history of this debate, see Church (2011); for an optimistic view of the potential symbiosis between computational linguistics and deep learning, see Manning (2015).

1.2.2 Search and learning

Many natural language processing problems can be written mathematically in the form of optimization,⁶

$$\hat{y} = \underset{y \in \mathcal{Y}(x)}{\operatorname{argmax}} \Psi(x, y; \theta), \quad [1.1]$$

where,

- x is the input, which is an element of a set \mathcal{X} ;
- y is the output, which is an element of a set $\mathcal{Y}(x)$;

⁵Ali Rahimi argued that much of deep learning research was similar to “alchemy” in a presentation at the 2017 conference on Neural Information Processing Systems. He was advocating for more learning theory, not more linguistics.

⁶Throughout this text, equations will be numbered by square brackets, and linguistic examples will be numbered by parentheses.

- Ψ is a scoring function (also called the **model**), which maps from the set $\mathcal{X} \times \mathcal{Y}$ to the real numbers;
- θ is a vector of parameters for Ψ ;
- \hat{y} is the predicted output, which is chosen to maximize the scoring function.

This basic structure can be applied to a huge range of problems. For example, the input x might be a social media post, and the output y might be a labeling of the emotional sentiment expressed by the author (chapter 4); or x could be a sentence in French, and the output y could be a sentence in Tamil (chapter 18); or x might be a sentence in English, and y might be a representation of the syntactic structure of the sentence (chapter 10); or x might be a news article and y might be a structured record of the events that the article describes (chapter 17).

This formulation reflects an implicit decision that language processing algorithms will have two distinct modules:

Search. The search module is responsible for computing the argmax of the function Ψ . In other words, it finds the output \hat{y} that gets the best score with respect to the input x . This is easy when the search space $\mathcal{Y}(x)$ is small enough to enumerate, or when the scoring function Ψ has a convenient decomposition into parts. In many cases, we will want to work with scoring functions that do not have these properties, motivating the use of more sophisticated search algorithms, such as bottom-up dynamic programming (§ 10.1) and beam search (§ 11.3.1). Because the outputs are usually discrete in language processing problems, search often relies on the machinery of **combinatorial optimization**.

Learning. The learning module is responsible for finding the parameters θ . This is typically (but not always) done by processing a large dataset of labeled examples, $\{(x^{(i)}, y^{(i)})\}_{i=1}^N$. Like search, learning is also approached through the framework of optimization, as we will see in chapter 2. Because the parameters are usually continuous, learning algorithms generally rely on **numerical optimization** to identify vectors of real-valued parameters that optimize some function of the model and the labeled data. Some basic principles of numerical optimization are reviewed in Appendix B.

The division of natural language processing into separate modules for search and learning makes it possible to reuse generic algorithms across many tasks and models. Much of the work of natural language processing can be focused on the design of the model Ψ — identifying and formalizing the linguistic phenomena that are relevant to the task at hand — while reaping the benefits of decades of progress in search, optimization, and learning. This textbook will describe several classes of scoring functions, and the corresponding algorithms for search and learning.

When a model is capable of making subtle linguistic distinctions, it is said to be *expressive*. Expressiveness is often traded off against efficiency of search and learning. For example, a word-to-word translation model makes search and learning easy, but it is not expressive enough to distinguish good translations from bad ones. Many of the most important problems in natural language processing seem to require expressive models, in which the complexity of search grows exponentially with the size of the input. In these models, exact search is usually impossible. Intractability threatens the neat modular composition between search and learning: if search requires a set of heuristic approximations, then it may be advantageous to learn a model that performs well under these specific heuristics. This has motivated some researchers to take a more integrated approach to search and learning, as briefly mentioned in chapters 11 and 15.

1.2.3 Relational, compositional, and distributional perspectives

Any element of language — a word, a phrase, a sentence, or even a sound — can be described from at least three perspectives. Consider the word *journalist*. A *journalist* is a subcategory of a *profession*, and an *anchorwoman* is a subcategory of *journalist*; furthermore, a *journalist* performs *journalism*, which is often, but not always, a subcategory of *writing*. This relational perspective on meaning is the basis for semantic **ontologies** such as WORDNET (Fellbaum, 2010), which enumerate the relations that hold between words and other elementary semantic units. The power of the relational perspective is illustrated by the following example:

(1.3) Umashanthi interviewed Ana. She works for the college newspaper.

Who works for the college newspaper? The word *journalist*, while not stated in the example, implicitly links the *interview* to the *newspaper*, making *Umashanthi* the most likely referent for the pronoun. (A general discussion of how to resolve pronouns is found in chapter 15.)

Yet despite the inferential power of the relational perspective, it is not easy to formalize computationally. Exactly which elements are to be related? Are *journalists* and *reporters* distinct, or should we group them into a single unit? Is the kind of *interview* performed by a journalist the same as the kind that one undergoes when applying for a job? Ontology designers face many such thorny questions, and the project of ontology design hearkens back to Borges' (1993) *Celestial Emporium of Benevolent Knowledge*, which divides animals into:

- (a) belonging to the emperor; (b) embalmed; (c) tame; (d) suckling pigs; (e) sirens; (f) fabulous; (g) stray dogs; (h) included in the present classification; (i) frenzied; (j) innumerable; (k) drawn with a very fine camelhair brush; (l) et cetera; (m) having just broken the water pitcher; (n) that from a long way off resemble flies.

735 Difficulties in ontology construction have led some linguists to argue that there is no task-
736 independent way to partition up word meanings (Kilgarriff, 1997).

737 Some problems are easier. Each member in a group of *journalists* is a *journalist*: the *-s*
738 suffix distinguishes the plural meaning from the singular in most of the nouns in English.
739 Similarly, a *journalist* can be thought of, perhaps colloquially, as someone who produces or
740 works on a *journal*. (Taking this approach even further, the word *journal* derives from the
741 French *jour+nal*, or *day+ly* = *daily*.) In this way, the meaning of a word is constructed from
742 the constituent parts — the principle of **compositionality**. This principle can be applied
743 to larger units: phrases, sentences, and beyond. Indeed, one of the great strengths of the
744 compositional view of meaning is that it provides a roadmap for understanding entire
745 texts and dialogues through a single analytic lens, grounding out in the smallest parts of
746 individual words.

747 But alongside *journalists* and *anti-parliamentarians*, there are many words that seem
748 to be linguistic atoms: think, for example, of *whale*, *blubber*, and *Nantucket*. Idiomatic
749 phrases like *kick the bucket* and *shoot the breeze* have meanings that are quite different from
750 the sum of their parts (Sag et al., 2002). Composition is of little help for such words and
751 expressions, but their meanings can be ascertained — or at least approximated — from the
752 contexts in which they appear. Take, for example, *blubber*, which appears in such contexts
753 as:

- 754 (1.4) a. The blubber served them as fuel.
755 b. ...extracting it from the blubber of the large fish ...
756 c. Amongst oily substances, blubber has been employed as a manure.

757 These contexts form the **distributional properties** of the word *blubber*, and they link it to
758 words which can appear in similar constructions: *fat*, *pelts*, and *barnacles*. This distribu-
759 tional perspective makes it possible to learn about meaning from unlabeled data alone;
760 unlike relational and compositional semantics, no manual annotation or expert knowl-
761 edge is required. Distributional semantics is thus capable of covering a huge range of
762 linguistic phenomena. However, it lacks precision: *blubber* is similar to *fat* in one sense, to
763 *pelts* in another sense, and to *barnacles* in still another. The question of *why* all these words
764 tend to appear in the same contexts is left unanswered.

765 The relational, compositional, and distributional perspectives all contribute to our un-
766 derstanding of linguistic meaning, and all three appear to be critical to natural language
767 processing. Yet they are uneasy collaborators, requiring seemingly incompatible represen-
768 tations and algorithmic approaches. This text presents some of the best known and most
769 successful methods for working with each of these representations, but future research
770 may reveal new ways to combine them.

771

Part I

772

Learning

773 **Chapter 2**

774 **Linear text classification**

775 We begin with the problem of **text classification**: given a text document, assign it a dis-
776 crete label $y \in \mathcal{Y}$, where \mathcal{Y} is the set of possible labels. Text classification has many appli-
777 cations, from spam filtering to analysis of electronic health records. This chapter describes
778 some of the most well known and effective algorithms for text classification, from a math-
779 ematical perspective that should help you understand what they do and why they work.
780 Text classification is also a building block that is used in more elaborate natural language
781 processing tasks. For readers without a background in machine learning or statistics, the
782 material in this chapter will take more time to digest than most of the subsequent chap-
783 ters. But this investment will pay off as the mathematical principles behind these basic
784 classification algorithms reappear in other contexts throughout the book.

785 **2.1 The bag of words**

786 To perform text classification, the first question is how to represent each document, or
787 instance. A common approach is to use a column vector of word counts, e.g., $\mathbf{x} =$
788 $[0, 1, 1, 0, 0, 2, 0, 1, 13, 0 \dots]^T$, where x_j is the count of word j . The length of \mathbf{x} is $V \triangleq |\mathcal{V}|$,
789 where \mathcal{V} is the set of possible words in the vocabulary. In linear classification, the classi-
790 fication decision is based on a weighted sum of individual feature counts, such as word
791 counts.

792 The object \mathbf{x} is a vector, but it is often called a **bag of words**, because it includes only
793 information about the count of each word, and not the order in which the words appear.
794 With the bag of words representation, we are ignoring grammar, sentence boundaries,
795 paragraphs — everything but the words. Yet the bag of words model is surprisingly
796 effective for text classification. If you see the word *whale* in a document, is it fiction or non-
797 fiction? What if you see the word *molybdenum*? For many labeling problems, individual
798 words can be strong predictors.

799 To predict a label from a bag-of-words, we can assign a score to each word in the vocabulary, measuring the compatibility with the label. For example, for the label FICTION,
 800 we might assign a positive score to the word *whale*, and a negative score to the word
 801 *molybdenum*. These scores are called **weights**, and they are arranged in a column vector θ .

803 Suppose that you want a multiclass classifier, where $K \triangleq |\mathcal{Y}| > 2$. For example, you
 804 might want to classify news stories about sports, celebrities, music, and business. The goal
 805 is to predict a label \hat{y} , given the bag of words x , using the weights θ . For each label $y \in \mathcal{Y}$,
 806 we compute a score $\Psi(x, y)$, which is a scalar measure of the compatibility between the
 807 bag-of-words x and the label y . In a linear bag-of-words classifier, this score is the vector
 808 inner product between the weights θ and the output of a **feature function** $f(x, y)$,

$$\Psi(x, y) = \theta \cdot f(x, y) = \sum_j \theta_j f_j(x, y). \quad [2.1]$$

809 As the notation suggests, f is a function of two arguments, the word counts x and the
 810 label y , and it returns a vector output. For example, given arguments x and y , element j
 811 of this feature vector might be,

$$f_j(x, y) = \begin{cases} x_{\text{whale}}, & \text{if } y = \text{FICTION} \\ 0, & \text{otherwise} \end{cases} \quad [2.2]$$

812 This function returns the count of the word *whale* if the label is FICTION, and it returns zero
 813 otherwise. The index j depends on the position of *whale* in the vocabulary, and of FICTION
 814 in the set of possible labels. The corresponding weight θ_j then scores the compatibility of
 815 the word *whale* with the label FICTION.¹ A positive score means that this word makes the
 816 label more likely.

The output of the feature function can be formalized as a vector:

$$f(x, y = 1) = [x; \underbrace{0; 0; \dots; 0}_{(K-1) \times V}] \quad [2.3]$$

$$f(x, y = 2) = [\underbrace{0; 0; \dots; 0}_V; x; \underbrace{0; 0; \dots; 0}_{(K-2) \times V}] \quad [2.4]$$

$$f(x, y = K) = [\underbrace{0; 0; \dots; 0}_{(K-1) \times V}; x], \quad [2.5]$$

817 where $\underbrace{[0; 0; \dots; 0]}_{(K-1) \times V}$ is a column vector of $(K - 1) \times V$ zeros, and the semicolon indicates
 818 vertical concatenation. For each of the K possible labels, the feature function returns a

¹In practice, both f and θ may be implemented as a dictionary rather than vectors, so that it is not necessary to explicitly identify j . In such an implementation, the tuple (*whale*, FICTION) acts as a key in both dictionaries; the values in f are feature counts, and the values in θ are weights.

819 vector that is mostly zeros, with a column vector of word counts \mathbf{x} inserted in a location
 820 that depends on the specific label y . This arrangement is shown in Figure 2.1. The notation
 821 may seem awkward at first, but it generalizes to an impressive range of learning settings,
 822 particularly **structure prediction**, which is the focus of Chapters 7-11.

Given a vector of weights, $\boldsymbol{\theta} \in \mathbb{R}^{V^K}$, we can now compute the score $\Psi(\mathbf{x}, y)$ by Equation 2.1. This inner product gives a scalar measure of the compatibility of the observation \mathbf{x} with label y .² For any document \mathbf{x} , we predict the label \hat{y} ,

$$\hat{y} = \operatorname{argmax}_{y \in \mathcal{Y}} \Psi(\mathbf{x}, y) \quad [2.6]$$

$$\Psi(\mathbf{x}, y) = \boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{x}, y). \quad [2.7]$$

823 This inner product notation gives a clean separation between the *data* (\mathbf{x} and y) and the
 824 *parameters* ($\boldsymbol{\theta}$).

825 While vector notation is used for presentation and analysis, in code the weights and
 826 feature vector can be implemented as dictionaries. The inner product can then be com-
 827 puted as a loop. In python:

```
828 def compute_score(x, y, weights):
829     total = 0
830     for feature, count in feature_function(x, y).items():
831         total += weights[feature] * count
832     return total
```

833 This representation is advantageous because it avoids storing and iterating over the many
 834 features whose counts are zero.

835 It is common to add an **offset feature** at the end of the vector of word counts \mathbf{x} , which
 836 is always 1. We then have to also add an extra zero to each of the zero vectors, to make the
 837 vector lengths match. This gives the entire feature vector $\mathbf{f}(\mathbf{x}, y)$ a length of $(V + 1) \times K$.
 838 The weight associated with this offset feature can be thought of as a bias for or against
 839 each label. For example, if we expect most emails to be spam, then the weight for the
 840 offset feature for $y = \text{SPAM}$ should be larger than the weight for the offset feature for
 841 $y = \text{NOT-SPAM}$.

Returning to the weights $\boldsymbol{\theta}$, where do they come from? One possibility is to set them by hand. If we wanted to distinguish, say, English from Spanish, we can use English and Spanish dictionaries, and set the weight to one for each word that appears in the

²Only $V \times (K - 1)$ features and weights are necessary. By stipulating that $\Psi(\mathbf{x}, y = K) = 0$ regardless of \mathbf{x} , it is possible to implement any classification rule that can be achieved with $V \times K$ features and weights. This is the approach taken in binary classification rules like $y = \text{Sign}(\boldsymbol{\beta} \cdot \mathbf{x} + a)$, where $\boldsymbol{\beta}$ is a vector of weights, a is an offset, and the label set is $\mathcal{Y} = \{-1, 1\}$. However, for multiclass classification, it is more concise to write $\boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{x}, y)$ for all $y \in \mathcal{Y}$.

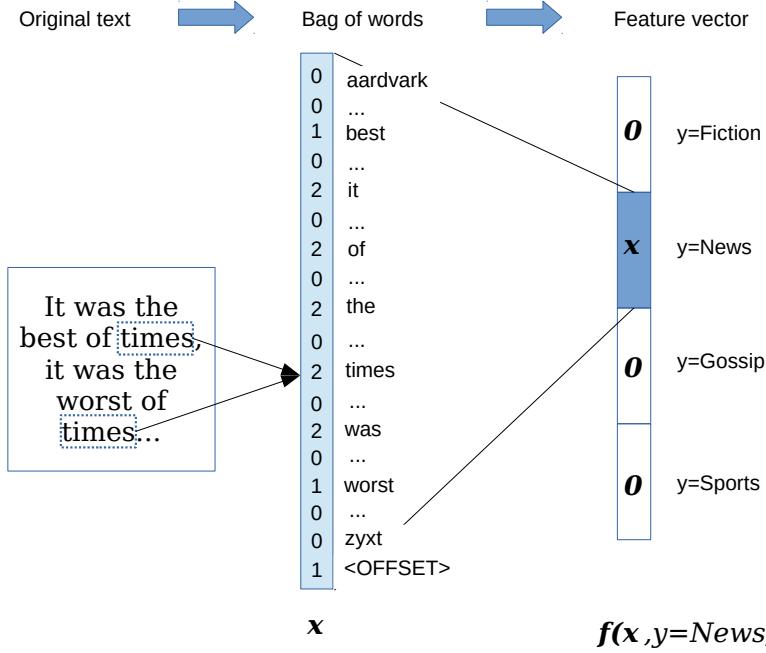


Figure 2.1: The bag-of-words and feature vector representations, for a hypothetical text classification task.

associated dictionary. For example,³

$$\begin{array}{ll}
 \theta_{(E,bicycle)} = 1 & \theta_{(S,bicycle)} = 0 \\
 \theta_{(E,bicicleta)} = 0 & \theta_{(S,bicicleta)} = 1 \\
 \theta_{(E,con)} = 1 & \theta_{(S,con)} = 1 \\
 \theta_{(E,ordinateur)} = 0 & \theta_{(S,ordinateur)} = 0.
 \end{array}$$

842 Similarly, if we want to distinguish positive and negative sentiment, we could use posi-
 843 tive and negative **sentiment lexicons** (see § 4.1.2), which are defined by social psycholo-
 844 gists (Tausczik and Pennebaker, 2010).

845 But it is usually not easy to set classification weights by hand, due to the large number
 846 of words and the difficulty of selecting exact numerical weights. Instead, we will learn the
 847 weights from data. Email users manually label messages as SPAM; newspapers label their
 848 own articles as BUSINESS or STYLE. Using such **instance labels**, we can automatically
 849 acquire weights using **supervised machine learning**. This chapter will discuss several
 850 machine learning approaches for classification. The first is based on probability. For a
 851 review of probability, consult Appendix A.

³In this notation, each tuple (language, word) indexes an element in θ , which remains a vector.

852 **2.2 Naïve Bayes**

853 The **joint probability** of a bag of words \mathbf{x} and its true label y is written $p(\mathbf{x}, y)$. Suppose
 854 we have a dataset of N labeled instances, $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^N$, which we assume are **independ-**
 855 **ent and identically distributed (IID)** (see § A.3). Then the joint probability of the entire
 856 dataset, written $p(\mathbf{x}^{(1:N)}, y^{(1:N)})$, is equal to $\prod_{i=1}^N p_{X,Y}(\mathbf{x}^{(i)}, y^{(i)})$.⁴

What does this have to do with classification? One approach to classification is to set the weights θ so as to maximize the joint probability of a **training set** of labeled documents. This is known as **maximum likelihood estimation**:

$$\hat{\theta} = \operatorname{argmax}_{\theta} p(\mathbf{x}^{(1:N)}, y^{(1:N)}; \theta) \quad [2.8]$$

$$= \operatorname{argmax}_{\theta} \prod_{i=1}^N p(\mathbf{x}^{(i)}, y^{(i)}; \theta) \quad [2.9]$$

$$= \operatorname{argmax}_{\theta} \sum_{i=1}^N \log p(\mathbf{x}^{(i)}, y^{(i)}; \theta). \quad [2.10]$$

857 The notation $p(\mathbf{x}^{(i)}, y^{(i)}; \theta)$ indicates that θ is a *parameter* of the probability function. The
 858 product of probabilities can be replaced by a sum of log-probabilities because the log func-
 859 tion is monotonically increasing over positive arguments, and so the same θ will maxi-
 860 mize both the probability and its logarithm. Working with logarithms is desirable because
 861 of numerical stability: on a large dataset, multiplying many probabilities can **underflow**
 862 to zero.⁵

863 The probability $p(\mathbf{x}^{(i)}, y^{(i)}; \theta)$ is defined through a **generative model** — an idealized
 864 random process that has generated the observed data.⁶ Algorithm 1 describes the gener-
 865 ative model underlying the **Naïve Bayes** classifier, with parameters $\theta = \{\mu, \phi\}$.

- 866 • The first line of this generative model encodes the assumption that the instances are
 867 mutually independent: neither the label nor the text of document i affects the label
 868 or text of document j .⁷ Furthermore, the instances are identically distributed: the

⁴The notation $p_{X,Y}(\mathbf{x}^{(i)}, y^{(i)})$ indicates the joint probability that random variables X and Y take the specific values $\mathbf{x}^{(i)}$ and $y^{(i)}$ respectively. The subscript will often be omitted when it is clear from context. For a review of random variables, see Appendix A.

⁵Throughout this text, you may assume all logarithms and exponents are base 2, unless otherwise indicated. Any reasonable base will yield an identical classifier, and base 2 is most convenient for working out examples by hand.

⁶Generative models will be used throughout this text. They explicitly define the assumptions underlying the form of a probability distribution over observed and latent variables. For a readable introduction to generative models in statistics, see Blei (2014).

⁷Can you think of any cases in which this assumption is too strong?

Algorithm 1 Generative process for the Naïve Bayes classification model

for Instance $i \in \{1, 2, \dots, N\}$ **do:**

 Draw the label $y^{(i)} \sim \text{Categorical}(\mu)$;

 Draw the word counts $\mathbf{x}^{(i)} | y^{(i)} \sim \text{Multinomial}(\phi_{y^{(i)}})$.

869 distributions over the label $y^{(i)}$ and the text $\mathbf{x}^{(i)}$ (conditioned on $y^{(i)}$) are the same
 870 for all instances i . In other words, we make the assumption that every document
 871 has the same distribution over labels, and that each document's distribution over
 872 words depends only on the label, and not on anything else about the document. We
 873 also assume that the documents don't affect each other: if the word *whale* appears
 874 in document $i = 7$, that does not make it any more or less likely that it will appear
 875 again in document $i = 8$.

- 876 • The second line of the generative model states that the random variable $y^{(i)}$ is drawn
 877 from a categorical distribution with parameter μ . Categorical distributions are like
 878 weighted dice: the column vector $\mu = [\mu_1; \mu_2; \dots; \mu_K]$ gives the probabilities of
 879 each label, so that the probability of drawing label y is equal to μ_y . For example, if
 880 $\mathcal{Y} = \{\text{POSITIVE}, \text{NEGATIVE}, \text{NEUTRAL}\}$, we might have $\mu = [0.1; 0.7; 0.2]$. We require
 881 $\sum_{y \in \mathcal{Y}} \mu_y = 1$ and $\mu_y \geq 0, \forall y \in \mathcal{Y}$: each label's probability is non-negative, and the
 882 sum of these probabilities is equal to one.⁸
- 883 • The third line describes how the bag-of-words counts $\mathbf{x}^{(i)}$ are generated. By writing
 884 $\mathbf{x}^{(i)} | y^{(i)}$, this line indicates that the word counts are conditioned on the label, so
 885 that the joint probability is factored using the chain rule,

$$p_{X,Y}(\mathbf{x}^{(i)}, y^{(i)}) = p_{X|Y}(\mathbf{x}^{(i)} | y^{(i)}) \times p_Y(y^{(i)}). \quad [2.11]$$

The specific distribution $p_{X|Y}$ is the **multinomial**, which is a probability distribution over vectors of non-negative counts. The probability mass function for this distribution is:

$$p_{\text{mult}}(\mathbf{x}; \phi) = B(\mathbf{x}) \prod_{j=1}^V \phi_j^{x_j} \quad [2.12]$$

$$B(\mathbf{x}) = \frac{(\sum_{j=1}^V x_j)!}{\prod_{j=1}^V (x_j!)}. \quad [2.13]$$

⁸Formally, we require $\mu \in \Delta^{K-1}$, where Δ^{K-1} is the $K - 1$ **probability simplex**, the set of all vectors of K nonnegative numbers that sum to one. Because of the sum-to-one constraint, there are $K - 1$ degrees of freedom for a vector of size K .

As in the categorical distribution, the parameter ϕ_j can be interpreted as a probability: specifically, the probability that any given token in the document is the word j . The multinomial distribution involves a product over words, with each term in the product equal to the probability ϕ_j , exponentiated by the count x_j . Words that have zero count play no role in this product, because $\phi_j^0 = 1$. The term $B(\mathbf{x})$ is called the **multinomial coefficient**. It doesn't depend on ϕ , and can usually be ignored. Can you see why we need this term at all?⁹

The notation $p(\mathbf{x} \mid y; \phi)$ indicates the conditional probability of word counts \mathbf{x} given label y , with parameter ϕ , which is equal to $p_{\text{mult}}(\mathbf{x}; \phi_y)$. By specifying the multinomial distribution, we describe the **multinomial Naïve Bayes** classifier. Why "naïve"? Because the multinomial distribution treats each word token independently, conditioned on the class: the probability mass function factorizes across the counts.¹⁰

2.2.1 Types and tokens

A slight modification to the generative model of Naïve Bayes is shown in Algorithm 2. Instead of generating a vector of counts of **types**, \mathbf{x} , this model generates a *sequence of tokens*, $\mathbf{w} = (w_1, w_2, \dots, w_M)$. The distinction between types and tokens is critical: $x_j \in \{0, 1, 2, \dots, M\}$ is the count of word type j in the vocabulary, e.g., the number of times the word *cannibal* appears; $w_m \in \mathcal{V}$ is the identity of token m in the document, e.g. $w_m = \textit{cannibal}$.

The probability of the sequence \mathbf{w} is a product of categorical probabilities. Algorithm 2 makes a conditional independence assumption: each token $w_m^{(i)}$ is independent of all other tokens $w_{n \neq m}^{(i)}$, conditioned on the label $y^{(i)}$. This is identical to the "naïve" independence assumption implied by the multinomial distribution, and as a result, the optimal parameters for this model are identical to those in multinomial Naïve Bayes. For any instance, the probability assigned by this model is proportional to the probability under multinomial Naïve Bayes. The constant of proportionality is the multinomial coefficient $B(\mathbf{x})$. Because $B(\mathbf{x}) \geq 1$, the probability for a vector of counts \mathbf{x} is at least as large as the probability for a list of words \mathbf{w} that induces the same counts: there can be many word sequences that correspond to a single vector of counts. For example, *man bites dog* and *dog bites man* correspond to an identical count vector, $\{\textit{bites} : 1, \textit{dog} : 1, \textit{man} : 1\}$, and $B(\mathbf{x})$ is equal to the total number of possible word orderings for count vector \mathbf{x} .

⁹Technically, a multinomial distribution requires a second parameter, the total number of word counts in \mathbf{x} . In the bag-of-words representation is equal to the number of words in the document. However, this parameter is irrelevant for classification.

¹⁰You can plug in any probability distribution to the generative story and it will still be Naïve Bayes, as long as you are making the "naïve" assumption that the features are conditionally independent, given the label. For example, a multivariate Gaussian with diagonal covariance is naïve in exactly the same sense.

Algorithm 2 Alternative generative process for the Naïve Bayes classification model

```

for Instance  $i \in \{1, 2, \dots, N\}$  do:
    Draw the label  $y^{(i)} \sim \text{Categorical}(\mu)$ ;
    for Token  $m \in \{1, 2, \dots, M_i\}$  do:
        Draw the token  $w_m^{(i)} \mid y^{(i)} \sim \text{Categorical}(\phi_{y^{(i)}})$ .

```

918 Sometimes it is useful to think of instances as counts of types, \mathbf{x} ; other times, it is
 919 better to think of them as sequences of tokens, \mathbf{w} . If the tokens are generated from a
 920 model that assumes conditional independence, then these two views lead to probability
 921 models that are identical, except for a scaling factor that does not depend on the label or
 922 the parameters.

923 **2.2.2 Prediction**

The Naïve Bayes prediction rule is to choose the label y which maximizes $\log p(\mathbf{x}, y; \boldsymbol{\mu}, \boldsymbol{\phi})$:

$$\hat{y} = \underset{y}{\operatorname{argmax}} \log p(\mathbf{x}, y; \boldsymbol{\mu}, \boldsymbol{\phi}) \quad [2.14]$$

$$= \underset{y}{\operatorname{argmax}} \log p(\mathbf{x} \mid y; \boldsymbol{\phi}) + \log p(y; \boldsymbol{\mu}) \quad [2.15]$$

Now we can plug in the probability distributions from the generative story.

$$\log p(\mathbf{x} \mid y; \boldsymbol{\phi}) + \log p(y; \boldsymbol{\mu}) = \log \left[B(\mathbf{x}) \prod_{j=1}^V \phi_{y,j}^{x_j} \right] + \log \mu_y \quad [2.16]$$

$$= \log B(\mathbf{x}) + \sum_{j=1}^V x_j \log \phi_{y,j} + \log \mu_y \quad [2.17]$$

$$= \log B(\mathbf{x}) + \boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{x}, y), \quad [2.18]$$

where

$$\boldsymbol{\theta} = [\boldsymbol{\theta}^{(1)}; \boldsymbol{\theta}^{(2)}; \dots; \boldsymbol{\theta}^{(K)}] \quad [2.19]$$

$$\boldsymbol{\theta}^{(y)} = [\log \phi_{y,1}; \log \phi_{y,2}; \dots; \log \phi_{y,V}; \log \mu_y] \quad [2.20]$$

924 The feature function $\mathbf{f}(\mathbf{x}, y)$ is a vector of V word counts and an offset, padded by
 925 zeros for the labels not equal to y (see Equations 2.3-2.5, and Figure 2.1). This construction
 926 ensures that the inner product $\boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{x}, y)$ only activates the features whose weights are
 927 in $\boldsymbol{\theta}^{(y)}$. These features and weights are all we need to compute the joint log-probability
 928 $\log p(\mathbf{x}, y)$ for each y . This is a key point: through this notation, we have converted the
 929 problem of computing the log-likelihood for a document-label pair (\mathbf{x}, y) into the compu-
 930 tation of a vector inner product.

931 **2.2.3 Estimation**

932 The parameters of the categorical and multinomial distributions have a simple interpretation:
 933 they are vectors of expected frequencies for each possible event. Based on this
 934 interpretation, it is tempting to set the parameters empirically,

$$\phi_{y,j} = \frac{\text{count}(y, j)}{\sum_{j'=1}^V \text{count}(y, j')} = \frac{\sum_{i:y^{(i)}=y} x_j^{(i)}}{\sum_{j'=1}^V \sum_{i:y^{(i)}=y} x_{j'}^{(i)}}, \quad [2.21]$$

935 where $\text{count}(y, j)$ refers to the count of word j in documents with label y .

936 Equation 2.21 defines the **relative frequency estimate** for ϕ . It can be justified as a
 937 **maximum likelihood estimate**: the estimate that maximizes the probability $p(\mathbf{x}^{(1:N)}, y^{(1:N)}; \theta)$.
 938 Based on the generative model in Algorithm 1, the log-likelihood is,

$$\mathcal{L}(\phi, \mu) = \sum_{i=1}^N \log p_{\text{mult}}(\mathbf{x}^{(i)}; \phi_{y^{(i)}}) + \log p_{\text{cat}}(y^{(i)}; \mu), \quad [2.22]$$

which is now written as a function \mathcal{L} of the parameters ϕ and μ . Let's continue to focus on the parameters ϕ . Since $p(y)$ is constant with respect to ϕ , we can drop it:

$$\mathcal{L}(\phi) = \sum_{i=1}^N \log p_{\text{mult}}(\mathbf{x}^{(i)}; \phi_{y^{(i)}}) = \sum_{i=1}^N \log B(\mathbf{x}^{(i)}) + \sum_{j=1}^V x_j^{(i)} \log \phi_{y^{(i)}, j}, \quad [2.23]$$

939 where $B(\mathbf{x}^{(i)})$ is constant with respect to ϕ .

940 Maximum-likelihood estimation chooses ϕ to maximize the log-likelihood \mathcal{L} . How-
 941 ever, the solution must obey the following constraints:

$$\sum_{j=1}^V \phi_{y,j} = 1 \quad \forall y \quad [2.24]$$

942 These constraints can be incorporated by adding a set of Lagrange multipliers to the ob-
 943 jective (see Appendix B for more details). To solve for each θ_y , we maximize the Lagrangian,

$$\ell(\phi_y) = \sum_{i:y^{(i)}=y} \sum_{j=1}^V x_j^{(i)} \log \phi_{y,j} - \lambda \left(\sum_{j=1}^V \phi_{y,j} - 1 \right). \quad [2.25]$$

Differentiating with respect to the parameter $\phi_{y,j}$ yields,

$$\frac{\partial \ell(\phi_y)}{\partial \phi_{y,j}} = \sum_{i:y^{(i)}=y} x_j^{(i)} / \phi_{y,j} - \lambda. \quad [2.26]$$

The solution is obtained by setting each element in this vector of derivatives equal to zero,

$$\lambda \phi_{y,j} = \sum_{i:y^{(i)}=y} x_j^{(i)} \quad [2.27]$$

$$\phi_{y,j} \propto \sum_{i:y^{(i)}=y} x_j^{(i)} = \sum_{i=1}^N \delta(y^{(i)} = y) x_j^{(i)} = \text{count}(y, j), \quad [2.28]$$

where $\delta(y^{(i)} = y)$ is a **delta function**, also sometimes called an indicator function, which returns one if $y^{(i)} = y$. The symbol \propto indicates that $\phi_{y,j}$ is **proportional to** the right-hand side of the equation.

Equation 2.28 shows three different notations for the same thing: a sum over the word counts for all documents i such that the label $y^{(i)} = y$. This gives a solution for each ϕ_y up to a constant of proportionality. Now recall the constraint $\sum_{j=1}^V \phi_{y,j} = 1$, which arises because ϕ_y represents a vector of probabilities for each word in the vocabulary. This constraint leads to an exact solution, which does not depend on λ :

$$\phi_{y,j} = \frac{\text{count}(y, j)}{\sum_{j'=1}^V \text{count}(y, j')} \quad [2.29]$$

This is equal to the relative frequency estimator from Equation 2.21. A similar derivation gives $\mu_y \propto \sum_{i=1}^N \delta(y^{(i)} = y)$.

2.2.4 Smoothing

With text data, there are likely to be pairs of labels and words that never appear in the training set, leaving $\phi_{y,j} = 0$. For example, the word *molybdenum* may have never yet appeared in a work of fiction. But choosing a value of $\phi_{\text{FICTION}, \text{molybdenum}} = 0$ would allow this single feature to completely veto a label, since $p(\text{FICTION} \mid \mathbf{x}) = 0$ if $x_{\text{molybdenum}} > 0$.

This is undesirable, because it imposes high **variance**: depending on what data happens to be in the training set, we could get vastly different classification rules. One solution is to **smooth** the probabilities, by adding a “pseudocount” of α to each count, and then normalizing.

$$\phi_{y,j} = \frac{\alpha + \text{count}(y, j)}{V\alpha + \sum_{j'=1}^V \text{count}(y, j')} \quad [2.30]$$

This is called **Laplace smoothing**.¹¹ The pseudocount α is a **hyperparameter**, because it controls the form of the log-likelihood function, which in turn drives the estimation of ϕ .

¹¹Laplace smoothing has a Bayesian justification, in which the generative model is extended to include ϕ as a random variable. The resulting distribution over ϕ depends on both the data (\mathbf{x} and y) and the **prior probability** $p(\phi; \alpha)$. The corresponding estimate of ϕ is called **maximum a posteriori**, or MAP. This is in contrast with maximum likelihood, which depends only on the data.

960 Smoothing reduces variance, but moves us away from the maximum likelihood esti-
961 mate: it imposes a **bias**. In this case, the bias points towards uniform probabilities. Ma-
962 chine learning theory shows that errors on heldout data can be attributed to the sum of
963 bias and variance (Mohri et al., 2012). In general, techniques for reducing variance often
964 increase the bias, leading to a **bias-variance tradeoff**.

- 965 • Unbiased classifiers may **overfit** the training data, yielding poor performance on
966 unseen data.
- 967 • But if the smoothing is too large, the resulting classifier can **underfit** instead. In the
968 limit of $\alpha \rightarrow \infty$, there is zero variance: you get the same classifier, regardless of the
969 data. However, the bias is likely to be large.

970 Similar issues arise throughout machine learning. Later in this chapter we will encounter
971 **regularization**, which controls the bias-variance tradeoff for logistic regression and large-
972 margin classifiers (§ 2.5.1); § 3.3.2 describes techniques for controlling variance in deep
973 learning; chapter 6 describes more elaborate methods for smoothing empirical probabili-
974 ties.

975 2.2.5 Setting hyperparameters

976 Returning to Naïve Bayes, how should we choose the best value of hyperparameters like
977 α ? Maximum likelihood will not work: the maximum likelihood estimate of α on the
978 training set will always be $\alpha = 0$. In many cases, what we really want is **accuracy**: the
979 number of correct predictions, divided by the total number of predictions. (Other mea-
980 sures of classification performance are discussed in § 4.4.) As we will see, it is hard to opti-
981 mize for accuracy directly. But for scalar hyperparameters like α , tuning can be performed
982 by a simple heuristic called **grid search**: try a set of values (e.g., $\alpha \in \{0.001, 0.01, 0.1, 1, 10\}$),
983 compute the accuracy for each value, and choose the setting that maximizes the accuracy.

984 The goal is to tune α so that the classifier performs well on *unseen* data. For this reason,
985 the data used for hyperparameter tuning should not overlap the training set, where very
986 small values of α will be preferred. Instead, we hold out a **development set** (also called
987 a **tuning set**) for hyperparameter selection. This development set may consist of a small
988 fraction of the labeled data, such as 10%.

989 We also want to predict the performance of our classifier on unseen data. To do this,
990 we must hold out a separate subset of data, called the **test set**. It is critical that the test set
991 not overlap with either the training or development sets, or else we will overestimate the
992 performance that the classifier will achieve on unlabeled data in the future. The test set
993 should also not be used when making modeling decisions, such as the form of the feature
994 function, the size of the vocabulary, and so on (these decisions are reviewed in chapter 4.)
995 The ideal practice is to use the test set only once — otherwise, the test set is used to guide

the classifier design, and test set accuracy will diverge from accuracy on truly unseen data. Because annotated data is expensive, this ideal can be hard to follow in practice, and many test sets have been used for decades. But in some high-impact applications like machine translation and information extraction, new test sets are released every year.

When only a small amount of labeled data is available, the test set accuracy can be unreliable. **K -fold cross-validation** is one way to cope with this scenario: the labeled data is divided into K folds, and each fold acts as the test set, while training on the other folds. The test set accuracies are then aggregated. In the extreme, each fold is a single data point; this is called **leave-one-out cross-validation**. To perform hyperparameter tuning in the context of cross-validation, another fold can be used for grid search. It is important not to repeatedly evaluate the cross-validated accuracy while making design decisions about the classifier, or you will overstate the accuracy on truly unseen data.

2.3 Discriminative learning

Naïve Bayes is easy to work with: the weights can be estimated in closed form, and the probabilistic interpretation makes it relatively easy to extend. However, the assumption that features are independent can seriously limit its accuracy. Thus far, we have defined the feature function $f(x, y)$ so that it corresponds to bag-of-words features: one feature per word in the vocabulary. In natural language, bag-of-words features violate the assumption of conditional independence — for example, the probability that a document will contain the word *naïve* is surely higher given that it also contains the word *Bayes* — but this violation is relatively mild.

However, good performance on text classification often requires features that are richer than the bag-of-words:

- To better handle out-of-vocabulary terms, we want features that apply to multiple words, such as prefixes and suffixes (e.g., *anti-*, *un-*, *-ing*) and capitalization.
- We also want n -gram features that apply to multi-word units: **bigrams** (e.g., *not good*, *not bad*), **trigrams** (e.g., *not so bad*, *lacking any decency*, *never before imagined*), and beyond.

These features flagrantly violate the Naïve Bayes independence assumption. Consider what happens if we add a prefix feature. Under the Naïve Bayes assumption, the joint probability of a word and its prefix are computed with the following approximation:¹²

$$\Pr(\text{word} = \textit{unfit}, \text{prefix} = \textit{un-} \mid y) \approx \Pr(\text{prefix} = \textit{un-} \mid y) \times \Pr(\text{word} = \textit{unfit} \mid y).$$

¹²The notation $\Pr(\cdot)$ refers to the probability of an event, and $p(\cdot)$ refers to the probability density or mass for a random variable (see Appendix A).

To test the quality of the approximation, we can manipulate the left-hand side by applying the chain rule,

$$\Pr(\text{word} = \text{unfit}, \text{prefix} = \text{un-} \mid y) = \Pr(\text{prefix} = \text{un-} \mid \text{word} = \text{unfit}, y) \quad [2.31]$$

$$\times \Pr(\text{word} = \text{unfit} \mid y) \quad [2.32]$$

But $\Pr(\text{prefix} = \text{un-} \mid \text{word} = \text{unfit}, y) = 1$, since *un-* is guaranteed to be the prefix for the word *unfit*. Therefore,

$$\Pr(\text{word} = \text{unfit}, \text{prefix} = \text{un-} \mid y) = 1 \times \Pr(\text{word} = \text{unfit} \mid y) \quad [2.33]$$

$$\gg \Pr(\text{prefix} = \text{un-} \mid y) \times \Pr(\text{word} = \text{unfit} \mid y), \quad [2.34]$$

because the probability of any given word starting with the prefix *un-* is much less than one. Naïve Bayes will systematically underestimate the true probabilities of conjunctions of positively correlated features. To use such features, we need learning algorithms that do not rely on an independence assumption.

The origin of the Naïve Bayes independence assumption is the learning objective, $p(\mathbf{x}^{(1:N)}, y^{(1:N)})$, which requires modeling the probability of the observed text. In classification problems, we are always given \mathbf{x} , and are only interested in predicting the label y . In this setting, modeling the probability of the text \mathbf{x} seems like a difficult and unnecessary task. **Discriminative learning** algorithms avoid this task, and focus directly on the problem of predicting y .

2.3.1 Perceptron

In Naïve Bayes, the weights can be interpreted as parameters of a probabilistic model. But this model requires an independence assumption that usually does not hold, and limits our choice of features. Why not forget about probability and learn the weights in an error-driven way? The **perceptron** algorithm, shown in Algorithm 3, is one way to do this.

The algorithm is simple: if you make a mistake, increase the weights for features that are active with the correct label $y^{(i)}$, and decrease the weights for features that are active with the guessed label \hat{y} . Perceptron is an **online learning** algorithm, since the classifier weights change after every example. This is different from Naïve Bayes, which is a **batch learning** algorithm: it computes statistics over the entire dataset, and then sets the weights in a single operation. Algorithm 3 is vague about when this online learning procedure terminates. We will return to this issue shortly.

The perceptron algorithm may seem like an unprincipled heuristic: Naïve Bayes has a solid foundation in probability, but the perceptron is just adding and subtracting constants from the weights every time there is a mistake. Will this really work? In fact, there is some nice theory for the perceptron, based on the concept of **linear separability**. Informally, a dataset with binary labels ($y \in \{0, 1\}$) is linearly separable if it is possible to draw a

Algorithm 3 Perceptron learning algorithm

```

1: procedure PERCEPTRON( $\mathbf{x}^{(1:N)}, y^{(1:N)}$ )
2:    $t \leftarrow 0$ 
3:    $\boldsymbol{\theta}^{(0)} \leftarrow \mathbf{0}$ 
4:   repeat
5:      $t \leftarrow t + 1$ 
6:     Select an instance  $i$ 
7:      $\hat{y} \leftarrow \operatorname{argmax}_y \boldsymbol{\theta}^{(t-1)} \cdot \mathbf{f}(\mathbf{x}^{(i)}, y)$ 
8:     if  $\hat{y} \neq y^{(i)}$  then
9:        $\boldsymbol{\theta}^{(t)} \leftarrow \boldsymbol{\theta}^{(t-1)} + \mathbf{f}(\mathbf{x}^{(i)}, y^{(i)}) - \mathbf{f}(\mathbf{x}^{(i)}, \hat{y})$ 
10:    else
11:       $\boldsymbol{\theta}^{(t)} \leftarrow \boldsymbol{\theta}^{(t-1)}$ 
12:    until tired
13:    return  $\boldsymbol{\theta}^{(t)}$ 

```

1051 hyperplane (a line in many dimensions), such that on each side of the hyperplane, all
 1052 instances have the same label. This definition can be formalized and extended to multiple
 1053 labels:

1054 **Definition 1** (Linear separability). *The dataset $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^N$ is linearly separable iff
 1055 (if and only if) there exists some weight vector $\boldsymbol{\theta}$ and some margin ρ such that for every instance
 1056 $(\mathbf{x}^{(i)}, y^{(i)})$, the inner product of $\boldsymbol{\theta}$ and the feature function for the true label, $\boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{x}^{(i)}, y^{(i)})$, is
 1057 at least ρ greater than inner product of $\boldsymbol{\theta}$ and the feature function for every other possible label,
 1058 $\boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{x}^{(i)}, y')$.*

$$\exists \boldsymbol{\theta}, \rho > 0 : \forall (\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{D}, \quad \boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{x}^{(i)}, y^{(i)}) \geq \rho + \max_{y' \neq y^{(i)}} \boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{x}^{(i)}, y'). \quad [2.35]$$

1059 Linear separability is important because of the following guarantee: if your data is
 1060 linearly separable, then the perceptron algorithm will find a separator (Novikoff, 1962).¹³
 1061 So while the perceptron may seem heuristic, it is guaranteed to succeed, if the learning
 1062 problem is easy enough.

1063 How useful is this proof? Minsky and Papert (1969) famously proved that the simple
 1064 logical function of *exclusive-or* is not separable, and that a perceptron is therefore inca-
 1065 pable of learning this function. But this is not just an issue for the perceptron: any linear
 1066 classification algorithm, including Naïve Bayes, will fail on this task. Text classification
 1067 problems usually involve high dimensional feature spaces, with thousands or millions of

¹³It is also possible to prove an upper bound on the number of training iterations required to find the separator. Proofs like this are part of the field of **machine learning theory** (Mohri et al., 2012).

1068 features. For these problems, it is very likely that the training data is indeed separable.
 1069 And even if the dataset is not separable, it is still possible to place an upper bound on the
 1070 number of errors that the perceptron algorithm will make (Freund and Schapire, 1999).

1071 2.3.2 Averaged perceptron

1072 The perceptron iterates over the data repeatedly — until “tired”, as described in Algo-
 1073 rithm 3. If the data is linearly separable, the perceptron will eventually find a separator,
 1074 and we can stop once all training instances are classified correctly. But if the data is not
 1075 linearly separable, the perceptron can *thrash* between two or more weight settings, never
 1076 converging. In this case, how do we know that we can stop training, and how should
 1077 we choose the final weights? An effective practical solution is to *average* the perceptron
 1078 weights across all iterations.

1079 This procedure is shown in Algorithm 4. The learning algorithm is nearly identical,
 1080 but we also maintain a vector of the sum of the weights, \mathbf{m} . At the end of the learning
 1081 procedure, we divide this sum by the total number of updates t , to compute the average
 1082 weights, $\bar{\theta}$. These average weights are then used for prediction. In the algorithm sketch,
 1083 the average is computed from a running sum, $\mathbf{m} \leftarrow \mathbf{m} + \theta$. However, this is inefficient,
 1084 because it requires $|\theta|$ operations to update the running sum. When $f(\mathbf{x}, y)$ is sparse,
 1085 $|\theta| \gg |f(\mathbf{x}, y)|$ for any individual (\mathbf{x}, y) . This means that computing the running sum will
 1086 be much more expensive than computing of the update to θ itself, which requires only
 1087 $2 \times |f(\mathbf{x}, y)|$ operations. One of the exercises is to sketch a more efficient algorithm for
 1088 computing the averaged weights.

1089 Even if the dataset is not separable, the averaged weights will eventually converge.
 1090 One possible stopping criterion is to check the difference between the average weight
 1091 vectors after each pass through the data: if the norm of the difference falls below some
 1092 predefined threshold, we can stop training. Another stopping criterion is to hold out some
 1093 data, and to measure the predictive accuracy on this heldout data. When the accuracy
 1094 on the heldout data starts to decrease, the learning algorithm has begun to **overfit** the
 1095 training set. At this point, it is probably best to stop; this stopping criterion is known as
 1096 **early stopping**.

1097 **Generalization** is the ability to make good predictions on instances that are not in
 1098 the training data. Averaging can be proven to improve generalization, by computing an
 1099 upper bound on the generalization error (Freund and Schapire, 1999; Collins, 2002).

1100 2.4 Loss functions and large-margin classification

1101 Naïve Bayes chooses the weights θ by maximizing the joint log-likelihood $\log p(\mathbf{x}^{(1:N)}, y^{(1:N)})$.
 1102 By convention, optimization problems are generally formulated as minimization of a **loss**
 1103 **function**. The input to a loss function is the vector of weights θ , and the output is a

Algorithm 4 Averaged perceptron learning algorithm

```

1: procedure AVG-PERCEPTRON( $\mathbf{x}^{(1:N)}, \mathbf{y}^{(1:N)}$ )
2:    $t \leftarrow 0$ 
3:    $\boldsymbol{\theta}^{(0)} \leftarrow 0$ 
4:   repeat
5:      $t \leftarrow t + 1$ 
6:     Select an instance  $i$ 
7:      $\hat{y} \leftarrow \operatorname{argmax}_y \boldsymbol{\theta}^{(t-1)} \cdot \mathbf{f}(\mathbf{x}^{(i)}, y)$ 
8:     if  $\hat{y} \neq y^{(i)}$  then
9:        $\boldsymbol{\theta}^{(t)} \leftarrow \boldsymbol{\theta}^{(t-1)} + \mathbf{f}(\mathbf{x}^{(i)}, y^{(i)}) - \mathbf{f}(\mathbf{x}^{(i)}, \hat{y})$ 
10:    else
11:       $\boldsymbol{\theta}^{(t)} \leftarrow \boldsymbol{\theta}^{(t-1)}$ 
12:     $\mathbf{m} \leftarrow \mathbf{m} + \boldsymbol{\theta}^{(t)}$ 
13:   until tired
14:    $\bar{\boldsymbol{\theta}} \leftarrow \frac{1}{t} \mathbf{m}$ 
15:   return  $\bar{\boldsymbol{\theta}}$ 

```

1104 non-negative number, measuring the performance of the classifier on a training instance.
 1105 Formally, the loss $\ell(\boldsymbol{\theta}; \mathbf{x}^{(i)}, y^{(i)})$ is then a measure of the performance of the weights $\boldsymbol{\theta}$ on
 1106 the instance $(\mathbf{x}^{(i)}, y^{(i)})$. The goal of learning is to minimize the sum of the losses across all
 1107 instances in the training set.

We can trivially reformulate maximum likelihood as a loss function, by defining the loss function to be the *negative log-likelihood*:

$$\log p(\mathbf{x}^{(1:N)}, \mathbf{y}^{(1:N)}; \boldsymbol{\theta}) = \sum_{i=1}^N \log p(\mathbf{x}^{(i)}, y^{(i)}; \boldsymbol{\theta}) \quad [2.36]$$

$$\ell_{\text{NB}}(\boldsymbol{\theta}; \mathbf{x}^{(i)}, y^{(i)}) = -\log p(\mathbf{x}^{(i)}, y^{(i)}; \boldsymbol{\theta}) \quad [2.37]$$

$$\hat{\boldsymbol{\theta}} = \operatorname{argmin}_{\boldsymbol{\theta}} \sum_{i=1}^N \ell_{\text{NB}}(\boldsymbol{\theta}; \mathbf{x}^{(i)}, y^{(i)}) \quad [2.38]$$

$$= \operatorname{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^N \log p(\mathbf{x}^{(i)}, y^{(i)}; \boldsymbol{\theta}). \quad [2.39]$$

1108 The problem of minimizing ℓ_{NB} is thus identical to maximum-likelihood estimation.

1109 Loss functions provide a general framework for comparing learning objectives. For
 1110 example, an alternative loss function is the **zero-one loss**,

$$\ell_{0-1}(\boldsymbol{\theta}; \mathbf{x}^{(i)}, y^{(i)}) = \begin{cases} 0, & y^{(i)} = \operatorname{argmax}_y \boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{x}^{(i)}, y) \\ 1, & \text{otherwise} \end{cases} \quad [2.40]$$

1111 The zero-one loss is zero if the instance is correctly classified, and one otherwise. The
 1112 sum of zero-one losses is proportional to the error rate of the classifier on the training
 1113 data. Since a low error rate is often the ultimate goal of classification, this may seem
 1114 ideal. But the zero-one loss has several problems. One is that it is **non-convex**,¹⁴ which
 1115 means that there is no guarantee that gradient-based optimization will be effective. A
 1116 more serious problem is that the derivatives are useless: the partial derivative with respect
 1117 to any parameter is zero everywhere, except at the points where $\theta \cdot f(\mathbf{x}^{(i)}, y) = \theta \cdot f(\mathbf{x}^{(i)}, \hat{y})$
 1118 for some \hat{y} . At those points, the loss is discontinuous, and the derivative is undefined.

1119 The perceptron optimizes a loss function that has better properties for learning:

$$\ell_{\text{PERCEPTRON}}(\boldsymbol{\theta}; \mathbf{x}^{(i)}, y^{(i)}) = \max_{y \in \mathcal{Y}} \boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{x}^{(i)}, y) - \boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{x}^{(i)}, y^{(i)}), \quad [2.41]$$

1120 When $\hat{y} = y^{(i)}$, the loss is zero; otherwise, it increases linearly with the gap between the
 1121 score for the predicted label \hat{y} and the score for the true label $y^{(i)}$. Plotting this loss against
 1122 the input $\max_{y \in \mathcal{Y}} \boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{x}^{(i)}, y) - \boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{x}^{(i)}, y^{(i)})$ gives a hinge shape, motivating the name
 1123 **hinge loss**.

1124 To see why this is the loss function optimized by the perceptron, take the derivative
 1125 with respect to $\boldsymbol{\theta}$,

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ell_{\text{PERCEPTRON}}(\boldsymbol{\theta}; \mathbf{x}^{(i)}, y^{(i)}) = \mathbf{f}(\mathbf{x}^{(i)}, \hat{y}) - \mathbf{f}(\mathbf{x}^{(i)}, y^{(i)}). \quad [2.42]$$

1126 At each instance, the perceptron algorithm takes a step of magnitude one in the opposite
 1127 direction of this **gradient**, $\nabla_{\boldsymbol{\theta}} \ell_{\text{PERCEPTRON}} = \frac{\partial}{\partial \boldsymbol{\theta}} \ell_{\text{PERCEPTRON}}(\boldsymbol{\theta}; \mathbf{x}^{(i)}, y^{(i)})$. As we will see in
 1128 § 2.6, this is an example of the optimization algorithm **stochastic gradient descent**, applied
 1129 to the objective in Equation 2.41.

1130 ***Breaking ties with subgradient descent** ¹⁵ Careful readers will notice the tacit assumption
 1131 that there is a unique \hat{y} that maximizes $\boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{x}^{(i)}, y)$. What if there are two or more
 1132 labels that maximize this function? Consider binary classification: if the maximizer is $y^{(i)}$,
 1133 then the gradient is zero, and so is the perceptron update; if the maximizer is $\hat{y} \neq y^{(i)}$,
 1134 then the update is the difference $\mathbf{f}(\mathbf{x}^{(i)}, y^{(i)}) - \mathbf{f}(\mathbf{x}^{(i)}, \hat{y})$. The underlying issue is that the
 1135 perceptron loss is not **smooth**, because the first derivative has a discontinuity at the hinge
 1136 point, where the score for the true label $y^{(i)}$ is equal to the score for some other label \hat{y} . At
 1137 this point, there is no unique gradient; rather, there is a set of **subgradients**. A vector \mathbf{v} is

¹⁴A function f is **convex** iff $\alpha f(x_i) + (1-\alpha)f(x_j) \geq f(\alpha x_i + (1-\alpha)x_j)$, for all $\alpha \in [0, 1]$ and for all x_i and x_j on the domain of the function. In words, any weighted average of the output of f applied to any two points is larger than the output of f when applied to the weighted average of the same two points. Convexity implies that any local minimum is also a global minimum, and there are many effective techniques for optimizing convex functions (Boyd and Vandenberghe, 2004). See Appendix B for a brief review.

¹⁵Throughout this text, advanced topics will be marked with an asterisk.

1138 a subgradient of the function g at \mathbf{u}_0 iff $g(\mathbf{u}) - g(\mathbf{u}_0) \geq \mathbf{v} \cdot (\mathbf{u} - \mathbf{u}_0)$ for all \mathbf{u} . Graphically,
 1139 this defines the set of hyperplanes that include $g(\mathbf{u}_0)$ and do not intersect g at any other
 1140 point. As we approach the hinge point from the left, the gradient is $\mathbf{f}(\mathbf{x}, \hat{y}) - \mathbf{f}(\mathbf{x}, y)$; as we
 1141 approach from the right, the gradient is $\mathbf{0}$. At the hinge point, the subgradients include all
 1142 vectors that are bounded by these two extremes. In subgradient descent, *any* subgradient
 1143 can be used (Bertsekas, 2012). Since both $\mathbf{0}$ and $\mathbf{f}(\mathbf{x}, \hat{y}) - \mathbf{f}(\mathbf{x}, y)$ are subgradients at the
 1144 hinge point, either one can be used in the perceptron update. This means that if multiple
 1145 labels maximize $\theta \cdot \mathbf{f}(\mathbf{x}^{(i)}, y)$, any of them can be used in the perceptron update.

1146 **Perceptron versus Naïve Bayes** The perceptron loss function has some pros and cons
 1147 with respect to the negative log-likelihood loss implied by Naïve Bayes.

- 1148 • Both ℓ_{NB} and $\ell_{\text{PERCEPTRON}}$ are convex, making them relatively easy to optimize. How-
 1149 ever, ℓ_{NB} can be optimized in closed form, while $\ell_{\text{PERCEPTRON}}$ requires iterating over
 1150 the dataset multiple times.
- 1151 • ℓ_{NB} can suffer *infinite* loss on a single example, since the logarithm of zero probability
 1152 is negative infinity. Naïve Bayes will therefore overemphasize some examples, and
 1153 underemphasize others.
- 1154 • The Naïve Bayes classifier assumes that the observed features are conditionally in-
 1155 dependent, given the label, and the performance of the classifier depends on the
 1156 extent to which this assumption holds. The perceptron requires no such assump-
 1157 tion.
- 1158 • $\ell_{\text{PERCEPTRON}}$ treats all correct answers equally. Even if θ only gives the correct answer
 1159 by a tiny margin, the loss is still zero.

1160 2.4.1 Online large margin classification

1161 This last comment suggests a potential problem with the perceptron. Suppose a test ex-
 1162 ample is very close to a training example, but not identical. If the classifier only gets the
 1163 correct answer on the training example by a small amount, then it may give a different
 1164 answer on the nearby test instance. To formalize this intuition, define the **margin** as,

$$\gamma(\theta; \mathbf{x}^{(i)}, y^{(i)}) = \theta \cdot \mathbf{f}(\mathbf{x}^{(i)}, y^{(i)}) - \max_{y \neq y^{(i)}} \theta \cdot \mathbf{f}(\mathbf{x}^{(i)}, y). \quad [2.43]$$

The margin represents the difference between the score for the correct label $y^{(i)}$, and the score for the highest-scoring incorrect label. The intuition behind **large margin clas-
 1165 sification** is that it is not enough to label the training data correctly — the correct label
 1166 should be separated from other labels by a comfortable margin. This idea can be encoded

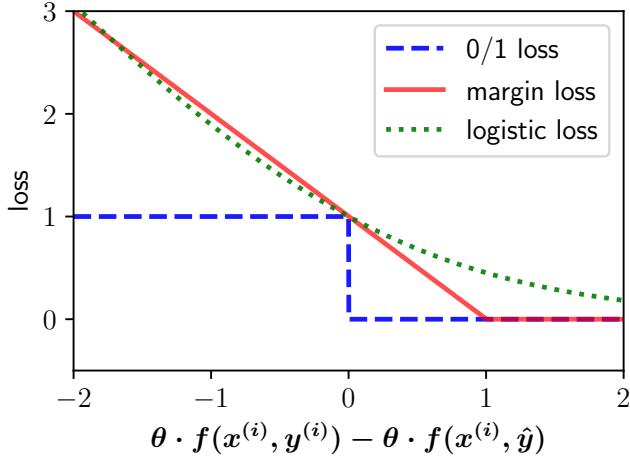


Figure 2.2: Margin, zero-one, and logistic loss functions.

into a loss function,

$$\ell_{\text{MARGIN}}(\boldsymbol{\theta}; \mathbf{x}^{(i)}, y^{(i)}) = \begin{cases} 0, & \gamma(\boldsymbol{\theta}; \mathbf{x}^{(i)}, y^{(i)}) \geq 1, \\ 1 - \gamma(\boldsymbol{\theta}; \mathbf{x}^{(i)}, y^{(i)}), & \text{otherwise} \end{cases} \quad [2.44]$$

$$= (1 - \gamma(\boldsymbol{\theta}; \mathbf{x}^{(i)}, y^{(i)}))_+, \quad [2.45]$$

where $(x)_+ = \max(0, x)$. The loss is zero if there is a margin of at least 1 between the score for the true label and the best-scoring alternative \hat{y} . This is almost identical to the perceptron loss, but the hinge point is shifted to the right, as shown in Figure 2.2. The margin loss is a convex upper bound on the zero-one loss.

The margin loss can be minimized using an online learning rule that is similar to perceptron. We will call this learning rule the **online support vector machine**, for reasons that will be discussed in the derivation. Let us first generalize the notion of a classification error with a **cost function** $c(y^{(i)}, y)$. We will focus on the simple cost function,

$$c(y^{(i)}, y) = \begin{cases} 1, & y^{(i)} \neq \hat{y} \\ 0, & \text{otherwise,} \end{cases} \quad [2.46]$$

but it is possible to design specialized cost functions that assign heavier penalties to especially undesirable errors (Tsochantaridis et al., 2004). This idea is revisited in chapter 7.

Using the cost function, we can now define the online support vector machine as the

following classification rule:

$$\hat{y} = \operatorname{argmax}_{y \in \mathcal{Y}} \theta \cdot f(\mathbf{x}^{(i)}, y) + c(y^{(i)}, y) \quad [2.47]$$

$$\theta^{(t)} \leftarrow (1 - \lambda)\theta^{(t-1)} + f(\mathbf{x}^{(i)}, y^{(i)}) - f(\mathbf{x}^{(i)}, \hat{y}) \quad [2.48]$$

1175 This update is similar in form to the perceptron, with two key differences.

- 1176 • Rather than selecting the label \hat{y} that maximizes the score of the current classifi-
1177 cation model, the argmax searches for labels that are both *strong*, as measured by
1178 $\theta \cdot f(\mathbf{x}^{(i)}, y)$, and *wrong*, as measured by $c(y^{(i)}, y)$. This maximization is known as
1179 **cost-augmented decoding**, because it augments the maximization objective to favor
1180 high-cost labels. If the highest-scoring label is $y = y^{(i)}$, then the margin loss for
1181 this instance is zero, and no update is needed. If not, then an update is required to
1182 reduce the margin loss — even if the current model classifies the instance correctly.
1183 Cost augmentation is only done while learning; it is not applied when making pre-
1184 dictions on unseen data.
- 1185 • The previous weights $\theta^{(t-1)}$ are scaled by $(1 - \lambda)$, with $\lambda \in (0, 1)$. The effect of this
1186 term is to cause the weights to “decay” back towards zero. In the support vector
1187 machine, this term arises from the minimization of a specific form of the margin, as
1188 described below. However, it can also be viewed as a form of **regularization**, which
1189 can help to prevent overfitting (see § 2.5.1). In this sense, it plays a role that is similar
1190 to smoothing in Naïve Bayes (see § 2.2.4).

1191 2.4.2 *Derivation of the online support vector machine

1192 The derivation of the online support vector machine is somewhat involved, but gives
1193 further intuition about why the method works. Begin by returning the idea of linear sep-
1194 arability (Definition 1): if a dataset is linearly separable, then there is some hyperplane θ
1195 that correctly classifies all training instances with margin ρ . This margin can be increased
1196 to any desired value by multiplying the weights by a constant.

Now, for any datapoint $(\mathbf{x}^{(i)}, y^{(i)})$, the geometric distance to the separating hyper-
plane is given by $\frac{\gamma(\theta; \mathbf{x}^{(i)}, y^{(i)})}{\|\theta\|_2}$, where the denominator is the norm of the weights, $\|\theta\|_2 = \sqrt{\sum_j \theta_j^2}$. The geometric distance is sometimes called the **geometric margin**, in contrast to
the **functional margin** $\gamma(\theta; \mathbf{x}^{(i)}, y^{(i)})$. Both are shown in Figure 2.3. The geometric margin
is a good measure of the robustness of the separator: if the functional margin is large, but
the norm $\|\theta\|_2$ is also large, then a small change in $\mathbf{x}^{(i)}$ could cause it to be misclassified.
We therefore seek to maximize the minimum geometric margin across the dataset, subject

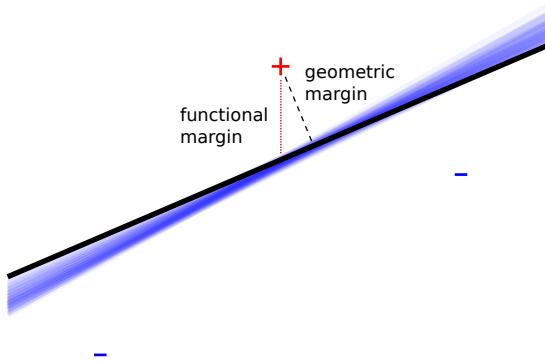


Figure 2.3: Functional and geometric margins for a binary classification problem. All separators that satisfy the margin constraint are shown. The separator with the largest geometric margin is shown in bold.

to the constraint that the margin loss is always zero:

$$\begin{aligned} \max_{\theta} \quad & \min_{i=1,2,\dots,N} \frac{\gamma(\theta; \mathbf{x}^{(i)}, y^{(i)})}{\|\theta\|_2} \\ \text{s.t.} \quad & \gamma(\theta; \mathbf{x}^{(i)}, y^{(i)}) \geq 1, \quad \forall i. \end{aligned} \quad [2.49]$$

1197 This is a **constrained optimization** problem, where the second line describes constraints
 1198 on the space of possible solutions θ . In this case, the constraint is that the functional
 1199 margin always be at least one, and the objective is that the minimum geometric margin
 1200 be as large as possible.

Constrained optimization is reviewed in Appendix B. In this case, further manipulation yields an unconstrained optimization problem. First, note that the norm $\|\theta\|_2$ scales linearly: $\|a\theta\|_2 = a\|\theta\|_2$. Furthermore, the functional margin γ is a linear function of θ , so that $\gamma(a\theta, \mathbf{x}^{(i)}, y^{(i)}) = a\gamma(\theta, \mathbf{x}^{(i)}, y^{(i)})$. As a result, any scaling factor on θ will cancel in the numerator and denominator of the geometric margin. If the data is linearly separable at any $\rho > 0$, it is always possible to rescale the functional margin to 1 by multiplying θ by a scalar constant. We therefore need only minimize the denominator $\|\theta\|_2$, subject to the constraint on the functional margin. The minimizer of $\|\theta\|_2$ is also the minimizer of $\frac{1}{2}\|\theta\|_2^2 = \frac{1}{2}\sum \theta_j^2$, which is easier to work with. This yields a simpler optimization prob-

lem:

$$\begin{aligned} \min_{\theta} . \quad & \frac{1}{2} \|\theta\|_2^2 \\ \text{s.t.} \quad & \gamma(\theta; \mathbf{x}^{(i)}, y^{(i)}) \geq 1, \quad \forall i. \end{aligned} \quad [2.50]$$

1201 This problem is a **quadratic program**: the objective is a quadratic function of the pa-
 1202 rameters, and the constraints are all linear inequalities. One solution to this problem is
 1203 to incorporate the constraints through Lagrange multipliers $\alpha_i \geq 0, i = 1, 2, \dots, N$. The
 1204 instances for which $\alpha_i > 0$ are called **support vectors**; other instances are irrelevant to the
 1205 classification boundary. This motivates the name **support vector machine**.

Thus far we have assumed linear separability, but many datasets of interest are not linearly separable. In this case, there is no θ that satisfies the margin constraint. To add more flexibility, we can introduce a set of **slack variables** $\xi_i \geq 0$. Instead of requiring that the functional margin be greater than or equal to one, we require that it be greater than or equal to $1 - \xi_i$. Ideally there would not be any slack, so the slack variables are penalized in the objective function:

$$\begin{aligned} \min_{\theta, \xi} \quad & \frac{1}{2} \|\theta\|_2^2 + C \sum_{i=1}^N \xi_i \\ \text{s.t.} \quad & \gamma(\theta; \mathbf{x}^{(i)}, y^{(i)}) + \xi_i \geq 1, \quad \forall i \\ & \xi_i \geq 0, \quad \forall i. \end{aligned} \quad [2.51]$$

1206 The hyperparameter C controls the tradeoff between violations of the margin con-
 1207 straint and the preference for a low norm of θ . As $C \rightarrow \infty$, slack is infinitely expensive,
 1208 and there is only a solution if the data is separable. As $C \rightarrow 0$, slack becomes free, and
 1209 there is a trivial solution at $\theta = 0$. Thus, C plays a similar role to the smoothing parame-
 1210 ter in Naïve Bayes (§ 2.2.4), trading off between a close fit to the training data and better
 1211 generalization. Like the smoothing parameter of Naïve Bayes, C must be set by the user,
 1212 typically by maximizing performance on a heldout development set.

1213 To solve the constrained optimization problem defined in Equation 2.51, we can first
 1214 solve for the slack variables,

$$\xi_i \geq (1 - \gamma(\theta; \mathbf{x}^{(i)}, y^{(i)}))_+. \quad [2.52]$$

The inequality is tight: the optimal solution is to make the slack variables as small as possible, while still satisfying the constraints (Ratliff et al., 2007; Smith, 2011). By plugging in the minimum slack variables back into Equation 2.51, the problem can be transformed into the unconstrained optimization,

$$\min_{\theta} \quad \frac{\lambda}{2} \|\theta\|_2^2 + \sum_{i=1}^N (1 - \gamma(\theta; \mathbf{x}^{(i)}, y^{(i)}))_+, \quad [2.53]$$

1215 where each ξ_i has been substituted by the right-hand side of Equation 2.52, and the factor
 1216 of C on the slack variables has been replaced by an equivalent factor of $\lambda = \frac{1}{C}$ on the
 1217 norm of the weights.

Equation 2.53 can be rewritten by expanding the margin,

$$\min_{\boldsymbol{\theta}} \quad \frac{\lambda}{2} \|\boldsymbol{\theta}\|_2^2 + \sum_{i=1}^N \left(\max_{y \in \mathcal{Y}} (\boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{x}^{(i)}, y) + c(y^{(i)}, y)) - \boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{x}^{(i)}, y^{(i)}) \right)_+, \quad [2.54]$$

where $c(y, y^{(i)})$ is the cost function defined in Equation 2.46. We can now differentiate with respect to the weights,

$$\nabla_{\boldsymbol{\theta}} L_{\text{SVM}} = \lambda \boldsymbol{\theta} + \sum_{i=1}^N \mathbf{f}(\mathbf{x}^{(i)}, \hat{y}) - \mathbf{f}(\mathbf{x}^{(i)}, y^{(i)}), \quad [2.55]$$

1218 where L_{SVM} refers to minimization objective in Equation 2.54 and $\hat{y} = \operatorname{argmax}_{y \in \mathcal{Y}} \boldsymbol{\theta} \cdot$
 1219 $\mathbf{f}(\mathbf{x}^{(i)}, y) + c(y^{(i)}, y)$. The online support vector machine update arises from the applica-
 1220 tion of **stochastic gradient descent** (described in § 2.6.2) to this gradient.

1221 2.5 Logistic regression

1222 Thus far, we have seen two broad classes of learning algorithms. Naïve Bayes is a prob-
 1223 abilistic method, where learning is equivalent to estimating a joint probability distribu-
 1224 tion. The perceptron and support vector machine are discriminative, error-driven algo-
 1225 rithms: the learning objective is closely related to the number of errors on the training
 1226 data. Probabilistic and error-driven approaches each have advantages: probability makes
 1227 it possible to quantify uncertainty about the predicted labels, but the probability model of
 1228 Naïve Bayes makes unrealistic independence assumptions that limit the features that can
 1229 be used.

Logistic regression combines advantages of discriminative and probabilistic classifiers. Unlike Naïve Bayes, which starts from the **joint probability** $p_{X,Y}$, logistic regression defines the desired **conditional probability** $p_{Y|X}$ directly. Think of $\boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{x}, y)$ as a scoring function for the compatibility of the base features \mathbf{x} and the label y . To convert this score into a probability, we first exponentiate, obtaining $\exp(\boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{x}, y))$, which is guaranteed to be non-negative. Next, we normalize, dividing over all possible labels $y' \in \mathcal{Y}$. The resulting conditional probability is defined as,

$$p(y | \mathbf{x}; \boldsymbol{\theta}) = \frac{\exp(\boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{x}, y))}{\sum_{y' \in \mathcal{Y}} \exp(\boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{x}, y'))}. \quad [2.56]$$

Given a dataset $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^N$, the weights θ are estimated by **maximum conditional likelihood**,

$$\log p(\mathbf{y}^{(1:N)} | \mathbf{x}^{(1:N)}; \theta) = \sum_{i=1}^N \log p(y^{(i)} | \mathbf{x}^{(i)}; \theta) \quad [2.57]$$

$$= \sum_{i=1}^N \theta \cdot \mathbf{f}(\mathbf{x}^{(i)}, y^{(i)}) - \log \sum_{y' \in \mathcal{Y}} \exp(\theta \cdot \mathbf{f}(\mathbf{x}^{(i)}, y')). \quad [2.58]$$

1230 The final line is obtained by plugging in Equation 2.56 and taking the logarithm.¹⁶ Inside
1231 the sum, we have the (additive inverse of the) **logistic loss**,

$$\ell_{\text{LOGREG}}(\theta; \mathbf{x}^{(i)}, y^{(i)}) = -\theta \cdot \mathbf{f}(\mathbf{x}^{(i)}, y^{(i)}) + \log \sum_{y' \in \mathcal{Y}} \exp(\theta \cdot \mathbf{f}(\mathbf{x}^{(i)}, y')) \quad [2.59]$$

1232 The logistic loss is shown in Figure 2.2 on page 31. A key difference from the zero-one
1233 and hinge losses is that logistic loss is never zero. This means that the objective function
1234 can always be improved by assigning higher confidence to the correct label.

1235 2.5.1 Regularization

1236 As with the support vector machine, better generalization can be obtained by penalizing
1237 the norm of θ . This is done by adding a multiple of the squared norm $\frac{\lambda}{2} \|\theta\|_2^2$ to the
1238 minimization objective. This is called L_2 regularization, because $\|\theta\|_2^2$ is the squared L_2
1239 norm of the vector θ . Regularization forces the estimator to trade off performance on the
1240 training data against the norm of the weights, and this can help to prevent overfitting.
1241 Consider what would happen to the unregularized weight for a base feature j that is
1242 active in only one instance $\mathbf{x}^{(i)}$: the conditional log-likelihood could always be improved
1243 by increasing the weight for this feature, so that $\theta_{(j,y^{(i)})} \rightarrow \infty$ and $\theta_{(j,\tilde{y} \neq y^{(i)})} \rightarrow -\infty$, where
1244 (j, y) is the index of feature associated with $x_j^{(i)}$ and label y in $\mathbf{f}(\mathbf{x}^{(i)}, y)$.

In § 2.2.4 (footnote 11), we saw that smoothing the probabilities of a Naïve Bayes classifier can be justified as a form of maximum a posteriori estimation, in which the parameters of the classifier are themselves random variables, drawn from a **prior distribution**. The same justification applies to L_2 regularization. In this case, the prior is a zero-mean Gaussian on each term of θ . The log-likelihood under a zero-mean Gaussian is,

$$\log N(\theta_j; 0, \sigma^2) \propto -\frac{1}{2\sigma^2} \theta_j^2, \quad [2.60]$$

1245 so that the regularization weight λ is equal to the inverse variance of the prior, $\lambda = \frac{1}{\sigma^2}$.

¹⁶The log-sum-exp term is a common pattern in machine learning. It is numerically unstable, because it will underflow if the inner product is small, and overflow if the inner product is large. Scientific computing libraries usually contain special functions for computing `logsumexp`, but with some thought, you should be able to see how to create an implementation that is numerically stable.

1246 **2.5.2 Gradients**

Logistic loss is minimized by optimization along the gradient. Specific algorithms are described in the next section, but first let's compute the gradient with respect to the logistic loss of a single example:

$$\ell_{\text{LOGREG}} = -\boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{x}^{(i)}, y^{(i)}) + \log \sum_{y' \in \mathcal{Y}} \exp(\boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{x}^{(i)}, y')) \quad [2.61]$$

$$\frac{\partial \ell}{\partial \boldsymbol{\theta}} = -\mathbf{f}(\mathbf{x}^{(i)}, y^{(i)}) + \frac{1}{\sum_{y'' \in \mathcal{Y}} \exp(\boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{x}^{(i)}, y''))} \times \sum_{y' \in \mathcal{Y}} \exp(\boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{x}^{(i)}, y')) \times \mathbf{f}(\mathbf{x}^{(i)}, y') \quad [2.62]$$

$$= -\mathbf{f}(\mathbf{x}^{(i)}, y^{(i)}) + \sum_{y' \in \mathcal{Y}} \frac{\exp(\boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{x}^{(i)}, y'))}{\sum_{y'' \in \mathcal{Y}} \exp(\boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{x}^{(i)}, y''))} \times \mathbf{f}(\mathbf{x}^{(i)}, y') \quad [2.63]$$

$$= -\mathbf{f}(\mathbf{x}^{(i)}, y^{(i)}) + \sum_{y' \in \mathcal{Y}} p(y' | \mathbf{x}^{(i)}; \boldsymbol{\theta}) \times \mathbf{f}(\mathbf{x}^{(i)}, y') \quad [2.64]$$

$$= -\mathbf{f}(\mathbf{x}^{(i)}, y^{(i)}) + E_{Y|X}[\mathbf{f}(\mathbf{x}^{(i)}, y)]. \quad [2.65]$$

1247 The final step employs the definition of a conditional expectation (§ A.5). The gradient of
 1248 the logistic loss is equal to the difference between the expected counts under the current
 1249 model, $E_{Y|X}[\mathbf{f}(\mathbf{x}^{(i)}, y)]$, and the observed feature counts $\mathbf{f}(\mathbf{x}^{(i)}, y^{(i)})$. When these two
 1250 vectors are equal for a single instance, there is nothing more to learn from it; when they
 1251 are equal in sum over the entire dataset, there is nothing more to learn from the dataset as
 1252 a whole. The gradient of the hinge loss is nearly identical, but it involves the features of
 1253 the predicted label under the current model, $\mathbf{f}(\mathbf{x}^{(i)}, \hat{y})$, rather than the expected features
 1254 $E_{Y|X}[\mathbf{f}(\mathbf{x}^{(i)}, y)]$ under the conditional distribution $p(y | \mathbf{x}; \boldsymbol{\theta})$.

The regularizer contributes $\lambda \boldsymbol{\theta}$ to the overall gradient:

$$L_{\text{LOGREG}} = \frac{\lambda}{2} \|\boldsymbol{\theta}\|_2^2 - \sum_{i=1}^N \left(\boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{x}^{(i)}, y^{(i)}) - \log \sum_{y' \in \mathcal{Y}} \exp \boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{x}^{(i)}, y') \right) \quad [2.66]$$

$$\nabla_{\boldsymbol{\theta}} L_{\text{LOGREG}} = \lambda \boldsymbol{\theta} - \sum_{i=1}^N \left(\mathbf{f}(\mathbf{x}^{(i)}, y^{(i)}) - E_{y|x}[\mathbf{f}(\mathbf{x}^{(i)}, y)] \right). \quad [2.67]$$

1255 **2.6 Optimization**

1256 Each of the classification algorithms in this chapter can be viewed as an optimization
 1257 problem:

- 1258 • In Naïve Bayes, the objective is the joint likelihood $\log p(\mathbf{x}^{(1:N)}, \mathbf{y}^{(1:N)})$. Maximum
 1259 likelihood estimation yields a closed-form solution for $\boldsymbol{\theta}$.

- 1260 • In the support vector machine, the objective is the regularized margin loss,

$$L_{\text{SVM}} = \frac{\lambda}{2} \|\boldsymbol{\theta}\|_2^2 + \sum_{i=1}^N (\max_{y \in \mathcal{Y}} (\boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{x}^{(i)}, y) + c(y^{(i)}, y)) - \boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{x}^{(i)}, y^{(i)}))_+, \quad [2.68]$$

1261 There is no closed-form solution, but the objective is convex. The perceptron algo-
1262 rithm minimizes a similar objective.

- 1263 • In logistic regression, the objective is the regularized negative log-likelihood,

$$L_{\text{LOGREG}} = \frac{\lambda}{2} \|\boldsymbol{\theta}\|_2^2 - \sum_{i=1}^N \left(\boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{x}^{(i)}, y^{(i)}) - \log \sum_{y \in \mathcal{Y}} \exp(\boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{x}^{(i)}, y)) \right) \quad [2.69]$$

1264 Again, there is no closed-form solution, but the objective is convex.

1265 These learning algorithms are distinguished by *what* is being optimized, rather than
1266 *how* the optimal weights are found. This decomposition is an essential feature of con-
1267 temporary machine learning. The domain expert's job is to design an objective function
1268 — or more generally, a **model** of the problem. If the model has certain characteristics,
1269 then generic optimization algorithms can be used to find the solution. In particular, if an
1270 objective function is differentiable, then gradient-based optimization can be employed;
1271 if it is also convex, then gradient-based optimization is guaranteed to find the globally
1272 optimal solution. The support vector machine and logistic regression have both of these
1273 properties, and so are amenable to generic **convex optimization** techniques (Boyd and
1274 Vandenberghe, 2004).

1275 2.6.1 Batch optimization

In **batch optimization**, each update to the weights is based on a computation involving the entire dataset. One such algorithm is **gradient descent**, which iteratively updates the weights,

$$\boldsymbol{\theta}^{(t+1)} \leftarrow \boldsymbol{\theta}^{(t)} - \eta^{(t)} \nabla_{\boldsymbol{\theta}} L, \quad [2.70]$$

1276 where $\nabla_{\boldsymbol{\theta}} L$ is the gradient computed over the entire training set, and $\eta^{(t)}$ is the **learning**
1277 **rate** at iteration t . If the objective L is a convex function of $\boldsymbol{\theta}$, then this procedure is
1278 guaranteed to terminate at the global optimum, for appropriate schedule of learning rates,
1279 $\eta^{(t)}$.¹⁷

¹⁷Convergence proofs typically require the learning rate to satisfy the following conditions: $\sum_{t=1}^{\infty} \eta^{(t)} = \infty$ and $\sum_{t=1}^{\infty} (\eta^{(t)})^2 < \infty$ (Bottou et al., 2016). These properties are satisfied by any learning rate schedule $\eta^{(t)} = \eta^{(0)} t^{-\alpha}$ for $\alpha \in [1, 2]$.

1280 In practice, gradient descent can be slow to converge, as the gradient can become
 1281 infinitesimally small. Faster convergence can be obtained by second-order Newton opti-
 1282 mization, which incorporates the inverse of the **Hessian matrix**,

$$H_{i,j} = \frac{\partial^2 L}{\partial \theta_i \partial \theta_j} \quad [2.71]$$

1283 The size of the Hessian matrix is quadratic in the number of features. In the bag-of-words
 1284 representation, this is usually too big to store, let alone invert. **Quasi-Network optimiza-**
 1285 **tion** techniques maintain a low-rank approximation to the inverse of the Hessian matrix.
 1286 Such techniques usually converge more quickly than gradient descent, while remaining
 1287 computationally tractable even for large feature sets. A popular quasi-Newton algorithm
 1288 is L-BFGS (Liu and Nocedal, 1989), which is implemented in many scientific computing
 1289 environments, such as SCIPY and MATLAB.

1290 For any gradient-based technique, the user must set the learning rates $\eta^{(t)}$. While con-
 1291 vergence proofs usually employ a decreasing learning rate, in practice, it is common to fix
 1292 $\eta^{(t)}$ to a small constant, like 10^{-3} . The specific constant can be chosen by experimentation,
 1293 although there is research on determining the learning rate automatically (Schaul et al.,
 1294 2013; Wu et al., 2018).

1295 2.6.2 Online optimization

1296 Batch optimization computes the objective on the entire training set before making an up-
 1297 date. This may be inefficient, because at early stages of training, a small number of train-
 1298 ing examples could point the learner in the correct direction. **Online learning** algorithms
 1299 make updates to the weights while iterating through the training data. The theoretical
 1300 basis for this approach is a stochastic approximation to the true objective function,

$$\sum_{i=1}^N \ell(\boldsymbol{\theta}; \mathbf{x}^{(i)}, y^{(i)}) \approx N \times \ell(\boldsymbol{\theta}; \mathbf{x}^{(j)}, y^{(j)}), \quad (\mathbf{x}^{(j)}, y^{(j)}) \sim \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^N, \quad [2.72]$$

1301 where the instance $(\mathbf{x}^{(j)}, y^{(j)})$ is sampled at random from the full dataset.

1302 In **stochastic gradient descent**, the approximate gradient is computed by randomly
 1303 sampling a single instance, and an update is made immediately. This is similar to the
 1304 perceptron algorithm, which also updates the weights one instance at a time. In **mini-**
 1305 **batch** stochastic gradient descent, the gradient is computed over a small set of instances.
 1306 A typical approach is to set the minibatch size so that the entire batch fits in memory on a
 1307 graphics processing unit (GPU; Neubig et al., 2017). It is then possible to speed up learn-
 1308 ing by parallelizing the computation of the gradient over each instance in the minibatch.

1309 Algorithm 5 offers a generalized view of gradient descent. In standard gradient de-
 1310 scent, the batcher returns a single batch with all the instances. In stochastic gradient de-

Algorithm 5 Generalized gradient descent. The function BATCHER partitions the training set into B batches such that each instance appears in exactly one batch. In gradient descent, $B = 1$; in stochastic gradient descent, $B = N$; in minibatch stochastic gradient descent, $1 < B < N$.

```

1: procedure GRADIENT-DESCENT( $\mathbf{x}^{(1:N)}, \mathbf{y}^{(1:N)}, L, \eta^{(1:\infty)}$ , BATCHER,  $T_{\max}$ )
2:    $\theta \leftarrow \mathbf{0}$ 
3:    $t \leftarrow 0$ 
4:   repeat
5:      $(\mathbf{b}^{(1)}, \mathbf{b}^{(2)}, \dots, \mathbf{b}^{(B)}) \leftarrow \text{BATCHER}(N)$ 
6:     for  $n \in \{1, 2, \dots, B\}$  do
7:        $t \leftarrow t + 1$ 
8:        $\theta^{(t)} \leftarrow \theta^{(t-1)} - \eta^{(t)} \nabla_{\theta} L(\theta^{(t-1)}; \mathbf{x}^{(b_1^{(n)}, b_2^{(n)}, \dots)}, \mathbf{y}^{(b_1^{(n)}, b_2^{(n)}, \dots)})$ 
9:       if Converged( $\theta^{(1,2,\dots,t)}$ ) then
10:        return  $\theta^{(t)}$ 
11:   until  $t \geq T_{\max}$ 
12:   return  $\theta^{(t)}$ 

```

1311 scent, it returns N batches with one instance each. In mini-batch settings, the batcher
 1312 returns B minibatches, $1 < B < N$.

There are many other techniques for online learning, and research in this area is ongoing (Bottou et al., 2016). Some algorithms use an adaptive learning rate, which can be different for every feature (Duchi et al., 2011). Features that occur frequently are likely to be updated frequently, so it is best to use a small learning rate; rare features will be updated infrequently, so it is better to take larger steps. The **AdaGrad** (adaptive gradient) algorithm achieves this behavior by storing the sum of the squares of the gradients for each feature, and rescaling the learning rate by its inverse:

$$\mathbf{g}_t = \nabla_{\theta} L(\theta^{(t)}; \mathbf{x}^{(i)}, y^{(i)}) \quad [2.73]$$

$$\theta_j^{(t+1)} \leftarrow \theta_j^{(t)} - \frac{\eta^{(t)}}{\sqrt{\sum_{t'=1}^t g_{t,j}^2}} g_{t,j}, \quad [2.74]$$

1313 where j iterates over features in $f(\mathbf{x}, y)$.

1314 In most cases, the number of active features for any instance is much smaller than the
 1315 number of weights. If so, the computation cost of online optimization will be dominated
 1316 by the update from the regularization term, $\lambda\theta$. The solution is to be “lazy”, updating
 1317 each θ_j only as it is used. To implement lazy updating, store an additional parameter τ_j ,
 1318 which is the iteration at which θ_j was last updated. If θ_j is needed at time t , the $t - \tau$
 1319 regularization updates can be performed all at once. This strategy is described in detail
 1320 by Kummerfeld et al. (2015).

1321 **2.7 *Additional topics in classification**

1322 This section presents some additional topics in classification that are particularly relevant
 1323 for natural language processing, especially for understanding the research literature.

1324 **2.7.1 Feature selection by regularization**

1325 In logistic regression and large-margin classification, generalization can be improved by
 1326 regularizing the weights towards 0, using the L_2 norm. But rather than encouraging
 1327 weights to be small, it might be better for the model to be **sparse**: it should assign weights
 1328 of exactly zero to most features, and only assign non-zero weights to features that are
 1329 clearly necessary. This idea can be formalized by the L_0 norm, $L_0 = \|\theta\|_0 = \sum_j \delta(\theta_j \neq 0)$,
 1330 which applies a constant penalty for each non-zero weight. This norm can be thought
 1331 of as a form of **feature selection**: optimizing the L_0 -regularized conditional likelihood is
 1332 equivalent to trading off the log-likelihood against the number of active features. Reduc-
 1333 ing the number of active features is desirable because the resulting model will be fast,
 1334 low-memory, and should generalize well, since irrelevant features will be pruned away.
 1335 Unfortunately, the L_0 norm is non-convex and non-differentiable. Optimization under L_0
 1336 regularization is **NP-hard**, meaning that it can be solved efficiently only if P=NP (Ge et al.,
 1337 2011).

1338 A useful alternative is the L_1 norm, which is equal to the sum of the absolute values
 1339 of the weights, $\|\theta\|_1 = \sum_j |\theta_j|$. The L_1 norm is convex, and can be used as an approxima-
 1340 tion to L_0 (Tibshirani, 1996). Conveniently, the L_1 norm also performs feature selection,
 1341 by driving many of the coefficients to zero; it is therefore known as a **sparsity inducing**
 1342 **regularizer**. The L_1 norm does not have a gradient at $\theta_j = 0$, so we must instead optimize
 1343 the L_1 -regularized objective using **subgradient** methods. The associated stochastic sub-
 1344 gradient descent algorithms are only somewhat more complex than conventional SGD;
 1345 Sra et al. (2012) survey approaches for estimation under L_1 and other regularizers.

1346 Gao et al. (2007) compare L_1 and L_2 regularization on a suite of NLP problems, finding
 1347 that L_1 regularization generally gives similar accuracy to L_2 regularization, but that L_1
 1348 regularization produces models that are between ten and fifty times smaller, because more
 1349 than 90% of the feature weights are set to zero.

1350 **2.7.2 Other views of logistic regression**

In binary classification, we can dispense with the feature function, and choose y based on the inner product of $\theta \cdot x$. The conditional probability $p_{Y|X}$ is obtained by passing this

inner product through a **logistic function**,

$$\sigma(a) \triangleq \frac{\exp(a)}{1 + \exp(a)} = (1 + \exp(-a))^{-1} \quad [2.75]$$

$$p(y | \mathbf{x}; \boldsymbol{\theta}) = \sigma(\boldsymbol{\theta} \cdot \mathbf{x}). \quad [2.76]$$

1351 This is the origin of the name “logistic regression.” Logistic regression can be viewed as
 1352 part of a larger family of **generalized linear models** (GLMs), in which various other **link**
 1353 **functions** convert between the inner product $\boldsymbol{\theta} \cdot \mathbf{x}$ and the parameter of a conditional
 1354 probability distribution.

1355 Logistic regression and related models are sometimes referred to as **log-linear**, be-
 1356 cause the log-probability is a linear function of the features. But in the early NLP liter-
 1357 ature, logistic regression was often called **maximum entropy** classification (Berger et al.,
 1358 1996). This name refers to an alternative formulation, in which the goal is to find the max-
 1359 imum entropy probability function that satisfies **moment-matching** constraints. These
 1360 constraints specify that the empirical counts of each feature should match the expected
 1361 counts under the induced probability distribution $p_{Y|X;\boldsymbol{\theta}}$,

$$\sum_{i=1}^N f_j(\mathbf{x}^{(i)}, y^{(i)}) = \sum_{i=1}^N \sum_{y \in \mathcal{Y}} p(y | \mathbf{x}^{(i)}; \boldsymbol{\theta}) f_j(\mathbf{x}^{(i)}, y), \quad \forall j \quad [2.77]$$

1362 The moment-matching constraint is satisfied exactly when the derivative of the condi-
 1363 tional log-likelihood function (Equation 2.65) is equal to zero. However, the constraint
 1364 can be met by many values of $\boldsymbol{\theta}$, so which should we choose?

1365 The **entropy** of the conditional probability distribution $p_{Y|X}$ is,

$$H(p_{Y|X}) = - \sum_{\mathbf{x} \in \mathcal{X}} p_X(\mathbf{x}) \sum_{y \in \mathcal{Y}} p_{Y|X}(y | \mathbf{x}) \log p_{Y|X}(y | \mathbf{x}), \quad [2.78]$$

1366 where \mathcal{X} is the set of all possible feature vectors, and $p_X(\mathbf{x})$ is the probability of observing
 1367 the base features \mathbf{x} . The distribution p_X is unknown, but it can be estimated by summing
 1368 over all the instances in the training set,

$$\tilde{H}(p_{Y|X}) = - \frac{1}{N} \sum_{i=1}^N \sum_{y \in \mathcal{Y}} p_{Y|X}(y | \mathbf{x}^{(i)}) \log p_{Y|X}(y | \mathbf{x}^{(i)}). \quad [2.79]$$

1369 If the entropy is large, the likelihood function is smooth across possible values of y ;
 1370 if it is small, the likelihood function is sharply peaked at some preferred value; in the
 1371 limiting case, the entropy is zero if $p(y | \mathbf{x}) = 1$ for some y . The maximum-entropy cri-
 1372 terion chooses to make the weakest commitments possible, while satisfying the moment-
 1373 matching constraints from Equation 2.77. The solution to this constrained optimization
 1374 problem is identical to the maximum conditional likelihood (logistic-loss) formulation
 1375 that was presented in § 2.5.

1376 2.8 Summary of learning algorithms

1377 It is natural to ask which learning algorithm is best, but the answer depends on what
1378 characteristics are important to the problem you are trying to solve.

1379 **Naïve Bayes** *Pros:* easy to implement; estimation is fast, requiring only a single pass over
1380 the data; assigns probabilities to predicted labels; controls overfitting with smoothing
1381 parameter. *Cons:* often has poor accuracy, especially with correlated features.

1382 **Perceptron** *Pros:* easy to implement; online; error-driven learning means that accuracy
1383 is typically high, especially after averaging. *Cons:* not probabilistic; hard to know
1384 when to stop learning; lack of margin can lead to overfitting.

1385 **Support vector machine** *Pros:* optimizes an error-based metric, usually resulting in high
1386 accuracy; overfitting is controlled by a regularization parameter. *Cons:* not probabilistic.
1387

1388 **Logistic regression** *Pros:* error-driven and probabilistic; overfitting is controlled by a regularization
1389 parameter. *Cons:* batch learning requires black-box optimization; logistic
1390 loss can “overtrain” on correctly labeled examples.

1391 One of the main distinctions is whether the learning algorithm offers a probability
1392 over labels. This is useful in modular architectures, where the output of one classifier
1393 is the input for some other system. In cases where probability is not necessary, the support
1394 vector machine is usually the right choice, since it is no more difficult to implement
1395 than the perceptron, and is often more accurate. When probability is necessary, logistic
1396 regression is usually more accurate than Naïve Bayes.

1397 Additional resources

1398 A machine learning textbook will offer more classifiers and more details (e.g., Murphy,
1399 2012), although the notation will differ slightly from what is typical in natural language
1400 processing. Probabilistic methods are surveyed by Hastie et al. (2009), and Mohri et al.
1401 (2012) emphasize theoretical considerations. Bottou et al. (2016) surveys the rapidly moving
1402 field of online learning, and Kummerfeld et al. (2015) empirically review several optimization
1403 algorithms for large-margin learning. The python toolkit SCIKIT-LEARN includes
1404 implementations of all of the algorithms described in this chapter (Pedregosa et al., 2011).

1405 Appendix B describes an alternative large-margin classifier, called **passive-aggressive**.
1406 Passive-aggressive is an online learner that seeks to make the smallest update that satisfies
1407 the margin constraint at the current instance. It is closely related to MIRA, which was used
1408 widely in NLP in the 2000s (Crammer and Singer, 2003).

1409 **Exercises**

1410 There will be exercises at the end of each chapter. In this chapter, the exercises are mostly
 1411 mathematical, matching the subject material. In other chapters, the exercises will emphasize
 1412 linguistics or programming.

- 1413 1. Let \mathbf{x} be a bag-of-words vector such that $\sum_{j=1}^V x_j = 1$. Verify that the multinomial
 1414 probability $p_{\text{mult}}(\mathbf{x}; \phi)$, as defined in Equation 2.12, is identical to the probability of
 1415 the same document under a categorical distribution, $p_{\text{cat}}(\mathbf{w}; \phi)$.
2. Suppose you have a single feature x , with the following conditional distribution:

$$p(x | y) = \begin{cases} \alpha, & X = 0, Y = 0 \\ 1 - \alpha, & X = 1, Y = 0 \\ 1 - \beta, & X = 0, Y = 1 \\ \beta, & X = 1, Y = 1. \end{cases} \quad [2.80]$$

1416 Further suppose that the prior is uniform, $\Pr(Y = 0) = \Pr(Y = 1) = \frac{1}{2}$, and that
 1417 both $\alpha > \frac{1}{2}$ and $\beta > \frac{1}{2}$. Given a Naïve Bayes classifier with accurate parameters,
 1418 what is the probability of making an error?

- 1419 3. Derive the maximum-likelihood estimate for the parameter μ in Naïve Bayes.
4. The classification models in the text have a vector of weights for each possible label.
 1420 While this is notationally convenient, it is overdetermined: for any linear classifier
 1421 that can be obtained with $K \times V$ weights, an equivalent classifier can be constructed
 1422 using $(K - 1) \times V$ weights.
 - 1424 a) Describe how to construct this classifier. Specifically, if given a set of weights
 1425 θ and a feature function $f(\mathbf{x}, y)$, explain how to construct alternative weights
 1426 and feature function θ' and $f'(\mathbf{x}, y)$, such that,

$$\forall y, y' \in \mathcal{Y}, \theta \cdot f(\mathbf{x}, y) - \theta \cdot f(\mathbf{x}, y') = \theta' \cdot f'(\mathbf{x}, y) - \theta' \cdot f'(\mathbf{x}, y'). \quad [2.81]$$

- 1427 b) Explain how your construction justifies the well-known alternative form for
 1428 binary logistic regression, $\Pr(Y = 1 | \mathbf{x}; \theta) = \frac{1}{1 + \exp(-\theta' \cdot \mathbf{x})} = \sigma(\theta' \cdot \mathbf{x})$, where σ
 1429 is the sigmoid function.
- 1430 5. Suppose you have two labeled datasets D_1 and D_2 , with the same features and la-
 1431 bels.
 - 1432 • Let $\theta^{(1)}$ be the unregularized logistic regression (LR) coefficients from training
 1433 on dataset D_1 .

- 1434 • Let $\theta^{(2)}$ be the unregularized LR coefficients (same model) from training on
 1435 dataset D_2 .
 1436 • Let θ^* be the unregularized LR coefficients from training on the combined
 1437 dataset $D_1 \cup D_2$.

Under these conditions, prove that for any feature j ,

$$\begin{aligned}\theta_j^* &\geq \min(\theta_j^{(1)}, \theta_j^{(2)}) \\ \theta_j^* &\leq \max(\theta_j^{(1)}, \theta_j^{(2)}).\end{aligned}$$

- 1438
- 1439 6. Let $\hat{\theta}$ be the solution to an unregularized logistic regression problem, and let θ^* be
 1440 the solution to the same problem, with L_2 regularization. Prove that $\|\theta^*\|_2^2 \leq \|\hat{\theta}\|_2^2$.
- 1441 7. As noted in the discussion of averaged perceptron in § 2.3.2, the computation of the
 1442 running sum $m \leftarrow m + \theta$ is unnecessarily expensive, requiring $K \times V$ operations.
 1443 Give an alternative way to compute the averaged weights $\bar{\theta}$, with complexity that is
 1444 independent of V and linear in the sum of feature sizes $\sum_{i=1}^N |\mathbf{f}(\mathbf{x}^{(i)}, y^{(i)})|$.
- 1445 8. Consider a dataset that is comprised of two identical instances $\mathbf{x}^{(1)} = \mathbf{x}^{(2)}$ with
 1446 distinct labels $y^{(1)} \neq y^{(2)}$. Assume all features are binary, $x_j \in \{0, 1\}$ for all j .

1447 Now suppose that the averaged perceptron always trains on the instance $(\mathbf{x}^{i(t)}, y^{i(t)})$,
 1448 where $i(t) = 2 - (t \bmod 2)$, which is 1 when the training iteration t is odd, and 2
 1449 when t is even. Further suppose that learning terminates under the following con-
 1450 dition:

$$\epsilon \geq \max_j \left| \frac{1}{t} \sum_t \theta_j^{(t)} - \frac{1}{t-1} \sum_t \theta_j^{(t-1)} \right|. \quad [2.82]$$

1451 In words, the algorithm stops when the largest change in the averaged weights is
 1452 less than or equal to ϵ . Compute the number of iterations before the averaged per-
 1453 ceptron terminates.

- 1454 9. Prove that the margin loss is convex in θ . Use this definition of the margin loss:

$$L(\theta) = -\theta \cdot \mathbf{f}(\mathbf{x}, y^*) + \max_y \theta \cdot \mathbf{f}(\mathbf{x}, y) + c(y^*, y), \quad [2.83]$$

1455 where y^* is the gold label. As a reminder, a function f is convex iff,

$$f(\alpha x_1 + (1 - \alpha)x_2) \leq \alpha f(x_1) + (1 - \alpha)f(x_2), \quad [2.84]$$

1456 for any x_1, x_2 and $\alpha \in [0, 1]$.

- 1457 10. If a function f is m -strongly convex, then for some $m > 0$, the following inequality
 1458 holds for all x and x' on the domain of the function:

$$f(x') \leq f(x) + (\nabla_x f) \cdot (x' - x) + \frac{m}{2} \|x' - x\|_2^2. \quad [2.85]$$

1459 Let $f(x) = L(\boldsymbol{\theta}^{(t)})$, representing the loss of the classifier at iteration t of gradient
 1460 descent; let $f(x') = L(\boldsymbol{\theta}^{(t+1)})$. Assuming the loss function is m -convex, prove that
 1461 $L(\boldsymbol{\theta}^{(t+1)}) \leq L(\boldsymbol{\theta}^{(t)})$ for an appropriate constant learning rate η , which will depend
 1462 on m . Explain why this implies that gradient descent converges when applied to an
 1463 m -strongly convex loss function with a unique minimum.

1464 **Chapter 3**

1465 **Nonlinear classification**

1466 Linear classification may seem like all we need for natural language processing. The bag-
1467 of-words representation is inherently high dimensional, and the number of features is
1468 often larger than the number of labeled training instances. This means that it is usually
1469 possible to find a linear classifier that perfectly fits the training data, or even to fit any ar-
1470bitrary labeling of the training instances! Moving to nonlinear classification may therefore
1471 only increase the risk of overfitting. Furthermore, for many tasks, **lexical features** (words)
1472 are meaningful in isolation, and can offer independent evidence about the instance label
1473 — unlike computer vision, where individual pixels are rarely informative, and must be
1474 evaluated holistically to make sense of an image. For these reasons, natural language
1475 processing has historically focused on linear classification.

1476 But in recent years, nonlinear classifiers have swept through natural language pro-
1477cessing, and are now the default approach for many tasks (Manning, 2015). There are at
1478 least three reasons for this change.

- 1479 • There have been rapid advances in **deep learning**, a family of nonlinear meth-
1480ods that learn complex functions of the input through multiple layers of computa-
1481tion (Goodfellow et al., 2016).
- 1482 • Deep learning facilitates the incorporation of **word embeddings**, which are dense
1483vector representations of words. Word embeddings can be learned from large amounts
1484of unlabeled data, and enable generalization to words that do not appear in the an-
1485notated training data (word embeddings are discussed in detail in chapter 14).
- 1486 • While CPU speeds have plateaued, there have been rapid advances in specialized
1487hardware called graphics processing units (GPUs), which have become faster, cheaper,
1488and easier to program. Many deep learning models can be implemented efficiently
1489on GPUs, offering substantial performance improvements over CPU-based comput-
1490ing.

1491 This chapter focuses on **neural networks**, which are the dominant approach for non-
 1492 linear classification in natural language processing today.¹ Historically, a few other non-
 1493 linear learning methods have been applied to language data.

- 1494 • **Kernel methods** are generalizations of the **nearest-neighbor** classification rule, which
 1495 classifies each instance by the label of the most similar example in the training set.
 1496 The application of the **kernel support vector machine** to information extraction is
 1497 described in chapter 17.
- 1498 • **Decision trees** classify instances by checking a set of conditions. Scaling decision
 1499 trees to bag-of-words inputs is difficult, but decision trees have been successful in
 1500 problems such as coreference resolution (chapter 15), where more compact feature
 1501 sets can be constructed (Soon et al., 2001).
- 1502 • **Boosting** and related **ensemble methods** work by combining the predictions of sev-
 1503 eral “weak” classifiers, each of which may consider only a small subset of features.
 1504 Boosting has been successfully applied to text classification (Schapire and Singer,
 1505 2000) and syntactic analysis (Abney et al., 1999), and remains one of the most suc-
 1506 cessful methods on machine learning competition sites such as Kaggle (Chen and
 1507 Guestrin, 2016).

1508 Hastie et al. (2009) provide an excellent overview of these techniques.

1509 3.1 Feedforward neural networks

1510 Consider the problem of building a classifier for movie reviews. The goal is to predict a
 1511 label $y \in \{\text{GOOD}, \text{BAD}, \text{OKAY}\}$ from a representation of the text of each document, x . But
 1512 what makes a good movie? The story, acting, cinematography, editing, soundtrack, and
 1513 so on. Now suppose the training set contains labels for each of these additional features,
 1514 $z = [z_1, z_2, \dots, z_{K_z}]^\top$. With a training set of such information, we could build a two-step
 1515 classifier:

- 1516 1. **Use the text x to predict the features z .** Specifically, train a logistic regression clas-
 1517 sifier to compute $p(z_k | x)$, for each $k \in \{1, 2, \dots, K_z\}$.
- 1518 2. **Use the features z to predict the label y .** Again, train a logistic regression classifier
 1519 to compute $p(y | z)$. On test data, z is unknown, so we will use the probabilities
 1520 $p(z | x)$ from the first layer as the features.

1521 This setup is shown in Figure 3.1, which describes the proposed classifier in a **compu-
 1522 tation graph**: the text features x are connected to the middle layer z , which is connected to
 1523 the label y .

¹I will use “deep learning” and “neural networks” interchangeably.

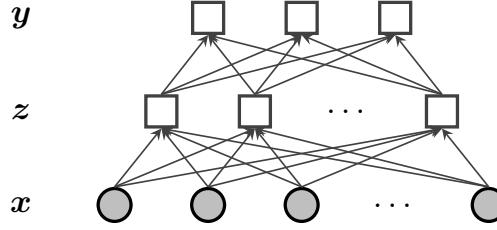


Figure 3.1: A feedforward neural network. Shaded circles indicate observed features, usually words; squares indicate nodes in the computation graph, which are computed from the information carried over the incoming arrows.

If we assume that each z_k is binary, $z_k \in \{0, 1\}$, then the probability $p(z_k | x)$ can be modeled using binary logistic regression:

$$\Pr(z_k = 1 | x; \Theta^{(x \rightarrow z)}) = \sigma(\theta_k^{(x \rightarrow z)} \cdot x) = (1 + \exp(-\theta_k^{(x \rightarrow z)} \cdot x))^{-1}, \quad [3.1]$$

where σ is the **sigmoid** function (shown in Figure 3.2), and the matrix $\Theta^{(x \rightarrow z)} \in \mathbb{R}^{K_z \times V}$ is constructed by stacking the weight vectors for each z_k ,

$$\Theta^{(x \rightarrow z)} = [\theta_1^{(x \rightarrow z)}, \theta_2^{(x \rightarrow z)}, \dots, \theta_{K_z}^{(x \rightarrow z)}]^\top. \quad [3.2]$$

We will assume that x contains a term with a constant value of 1, so that a corresponding offset parameter is included in each $\theta_k^{(x \rightarrow z)}$.

The output layer is computed by the multi-class logistic regression probability,

$$\Pr(y = j | z; \Theta^{(z \rightarrow y)}, b) = \frac{\exp(\theta_j^{(z \rightarrow y)} \cdot z + b_j)}{\sum_{j' \in \mathcal{Y}} \exp(\theta_{j'}^{(z \rightarrow y)} \cdot z + b_{j'})}, \quad [3.3]$$

where b_j is an offset for label j , and the output weight matrix $\Theta^{(z \rightarrow y)} \in \mathbb{R}^{K_y \times K_z}$ is again constructed by concatenation,

$$\Theta^{(z \rightarrow y)} = [\theta_1^{(z \rightarrow y)}, \theta_2^{(z \rightarrow y)}, \dots, \theta_{K_y}^{(z \rightarrow y)}]^\top. \quad [3.4]$$

The vector of probabilities over each possible value of y is denoted,

$$p(y | z; \Theta^{(z \rightarrow y)}, b) = \text{SoftMax}(\Theta^{(z \rightarrow y)} z + b), \quad [3.5]$$

where element j in the output of the **SoftMax** function is computed as in Equation 3.3.

This set of equations defines a multilayer classifier, which can be summarized as,

$$p(z | x; \Theta^{(x \rightarrow z)}) = \sigma(\Theta^{(x \rightarrow z)} x) \quad [3.6]$$

$$p(y | z; \Theta^{(z \rightarrow y)}, b) = \text{SoftMax}(\Theta^{(z \rightarrow y)} z + b), \quad [3.7]$$

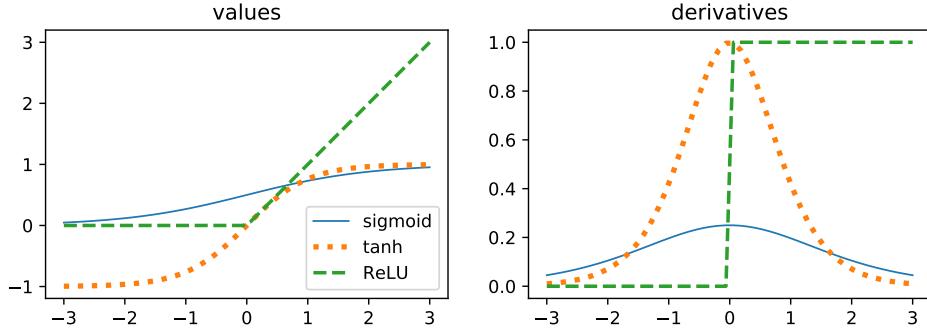


Figure 3.2: The sigmoid, tanh, and ReLU activation functions

1535 where the function σ is now applied **elementwise** to the vector of inner products,

$$\sigma(\Theta^{(x \rightarrow z)} \mathbf{x}) = [\sigma(\theta_1^{(x \rightarrow z)} \cdot \mathbf{x}), \sigma(\theta_2^{(x \rightarrow z)} \cdot \mathbf{x}), \dots, \sigma(\theta_{K_z}^{(x \rightarrow z)} \cdot \mathbf{x})]^\top. \quad [3.8]$$

Now suppose that the hidden features z are never observed, even in the training data. We can still construct the architecture in Figure 3.1. Instead of predicting y from a discrete vector of predicted values z , we use the probabilities $\sigma(\theta_k \cdot \mathbf{x})$. The resulting classifier is barely changed:

$$\mathbf{z} = \sigma(\Theta^{(x \rightarrow z)} \mathbf{x}) \quad [3.9]$$

$$p(y | \mathbf{x}; \Theta^{(z \rightarrow y)}, \mathbf{b}) = \text{SoftMax}(\Theta^{(z \rightarrow y)} \mathbf{z} + \mathbf{b}). \quad [3.10]$$

1536 This defines a classification model that predicts the label $y \in \mathcal{Y}$ from the base features \mathbf{x} ,
1537 through a “hidden layer” \mathbf{z} . This is a **feedforward neural network**.²

1538 3.2 Designing neural networks

1539 There several ways to generalize the feedforward neural network.

1540 3.2.1 Activation functions

1541 If the hidden layer is viewed as a set of latent features, then the sigmoid function in Equa-
1542 tion 3.9 represents the extent to which each of these features is “activated” by a given
1543 input. However, the hidden layer can be regarded more generally as a nonlinear trans-
1544 formation of the input. This opens the door to many other activation functions, some of
1545 which are shown in Figure 3.2. At the moment, the choice of activation functions is more
1546 art than science, but a few points can be made about the most popular varieties:

²The architecture is sometimes called a **multilayer perceptron**, but this is misleading, because each layer is not a perceptron as defined in the previous chapter.

- The range of the sigmoid function is $(0, 1)$. The bounded range ensures that a cascade of sigmoid functions will not “blow up” to a huge output, and this is important for deep networks with several hidden layers. The derivative of the sigmoid is $\frac{\partial}{\partial a}\sigma(a) = \sigma(a)(1 - \sigma(a))$. This derivative becomes small at the extremes, which can make learning slow; this is called the **vanishing gradient** problem.
- The range of the **tanh activation function** is $(-1, 1)$: like the sigmoid, the range is bounded, but unlike the sigmoid, it includes negative values. The derivative is $\frac{\partial}{\partial a}\tanh(a) = 1 - \tanh(a)^2$, which is steeper than the logistic function near the origin (LeCun et al., 1998). The tanh function can also suffer from vanishing gradients at extreme values.
- The **rectified linear unit (ReLU)** is zero for negative inputs, and linear for positive inputs (Glorot et al., 2011),

$$\text{ReLU}(a) = \begin{cases} a, & a \geq 0 \\ 0, & \text{otherwise.} \end{cases} \quad [3.11]$$

The derivative is a step function, which is 1 if the input is positive, and zero otherwise. Once the activation is zero, the gradient is also zero. This can lead to the problem of “dead neurons”, where some ReLU nodes are zero for all inputs, throughout learning. A solution is the **leaky ReLU**, which has a small positive slope for negative inputs (Maas et al., 2013),

$$\text{Leaky-ReLU}(a) = \begin{cases} a, & a \geq 0 \\ .0001a, & \text{otherwise.} \end{cases} \quad [3.12]$$

Sigmoid and tanh are sometimes described as **squashing functions**, because they squash an unbounded input into a bounded range. Glorot and Bengio (2010) recommend against the use of the sigmoid activation in deep networks, because its mean value of $\frac{1}{2}$ can cause the next layer of the network to be saturated, leading to small gradients on its own parameters. Several other activation functions are reviewed in the textbook by Goodfellow et al. (2016), who recommend ReLU as the “default option.”

3.2.2 Network structure

Deep networks stack up several hidden layers, with each $z^{(d)}$ acting as the input to the next layer, $z^{(d+1)}$. As the total number of nodes in the network increases, so does its capacity to learn complex functions of the input. Given a fixed number of nodes, one must decide whether to emphasize width (large K_z at each layer) or depth (many layers). At present, this tradeoff is not well understood.³

³With even a single hidden layer, a neural network can approximate any continuous function on a closed and bounded subset of \mathbb{R}^N to an arbitrarily small non-zero error; see section 6.4.1 of Goodfellow et al. (2016)

1576 It is also possible to “short circuit” a hidden layer, by propagating information directly
 1577 from the input to the next higher level of the network. This is the idea behind **residual net-**
 1578 **works**, which propagate information directly from the input to the subsequent layer (He
 1579 et al., 2016),

$$\mathbf{z} = f(\Theta^{(x \rightarrow z)} \mathbf{x}) + \mathbf{x}, \quad [3.13]$$

where f is any nonlinearity, such as sigmoid or ReLU. A more complex architecture is the **highway network** (Srivastava et al., 2015; Kim et al., 2016), in which an addition **gate** controls an interpolation between $f(\Theta^{(x \rightarrow z)} \mathbf{x})$ and \mathbf{x} ,

$$\mathbf{t} = \sigma(\Theta^{(t)} \mathbf{x} + \mathbf{b}^{(t)}) \quad [3.14]$$

$$\mathbf{z} = \mathbf{t} \odot f(\Theta^{(x \rightarrow z)} \mathbf{x}) + (\mathbf{1} - \mathbf{t}) \odot \mathbf{x}, \quad [3.15]$$

1580 where \odot refers to an elementwise vector product, and $\mathbf{1}$ is a column vector of ones. As
 1581 before, the sigmoid function is applied elementwise to its input; recall that the output of
 1582 this function is restricted to the range $(0, 1)$. Gating is also used in the **long short-term**
 1583 **memory (LSTM)**, which is discussed in chapter 6. Residual and highway connections
 1584 address a problem with deep architectures: repeated application of a nonlinear activation
 1585 function can make it difficult to learn the parameters of the lower levels of the network,
 1586 which are too distant from the supervision signal.

1587 3.2.3 Outputs and loss functions

In the multi-class classification example, a softmax output produces probabilities over each possible label. This aligns with a negative **conditional log-likelihood**,

$$-\mathcal{L} = -\sum_{i=1}^N \log p(y^{(i)} | \mathbf{x}^{(i)}; \Theta). \quad [3.16]$$

1588 where $\Theta = \{\Theta^{(x \rightarrow z)}, \Theta^{(z \rightarrow y)}, \mathbf{b}\}$ is the entire set of parameters.

This loss can be written alternatively as follows:

$$\tilde{y}_j \triangleq \Pr(y = j | \mathbf{x}^{(i)}; \Theta) \quad [3.17]$$

$$-\mathcal{L} = -\sum_{i=1}^N e_{y^{(i)}} \cdot \log \tilde{y} \quad [3.18]$$

for a survey of these theoretical results. However, depending on the function to be approximated, the width of the hidden layer may need to be arbitrarily large. Furthermore, the fact that a network has the *capacity* to approximate any given function does not imply that it is possible to *learn* the function using gradient-based optimization.

1589 where $e_{y^{(i)}}$ is a **one-hot vector** of zeros with a value of 1 at position $y^{(i)}$. The inner product
 1590 between $e_{y^{(i)}}$ and \tilde{y} is also called the multinomial **cross-entropy**, and this terminology
 1591 is preferred in many neural networks papers and software packages.

It is also possible to train neural networks from other objectives, such as a margin loss.
 In this case, it is not necessary to use softmax at the output layer: an affine transformation
 of the hidden layer is enough:

$$\Psi(y; \mathbf{x}^{(i)}, \Theta) = \theta_y^{(z \rightarrow y)} \cdot z + b_y \quad [3.19]$$

$$\ell_{\text{MARGIN}}(\Theta; \mathbf{x}^{(i)}, y^{(i)}) = \max_{y \neq y^{(i)}} \left(1 + \Psi(y; \mathbf{x}^{(i)}, \Theta) - \Psi(y^{(i)}; \mathbf{x}^{(i)}, \Theta) \right)_+ \quad [3.20]$$

1592 In regression problems, the output is a scalar or vector (see § 4.1.2). For these problems, a
 1593 typical loss function is the squared error $(y - \hat{y})^2$ or squared norm $\|\mathbf{y} - \hat{\mathbf{y}}\|_2^2$.

1594 3.2.4 Inputs and lookup layers

1595 In text classification, the input layer \mathbf{x} can refer to a bag-of-words vector, where x_j is
 1596 the count of word j . The input to the hidden unit z_k is then $\sum_{j=1}^V \theta_{j,k}^{(x \rightarrow z)} x_j$, and word j is
 1597 represented by the vector $\theta_j^{(x \rightarrow z)}$. This vector is sometimes described as the **embedding** of
 1598 word j , and can be learned from unlabeled data, using techniques discussed in chapter 14.
 1599 The columns of $\Theta^{(x \rightarrow z)}$ are each K_z -dimensional word embeddings.

1600 Chapter 2 presented an alternative view of text documents, as a sequence of word
 1601 tokens, w_1, w_2, \dots, w_M . In a neural network, each word token w_m is represented with a
 1602 one-hot vector, e_{w_m} , with dimension V . The matrix-vector product $\Theta^{(x \rightarrow z)} e_{w_m}$ returns
 1603 the embedding of word w_m . The complete document can be represented by horizontally
 1604 concatenating these one-hot vectors, $\mathbf{W} = [e_{w_1}, e_{w_2}, \dots, e_{w_M}]$, and the bag-of-words rep-
 1605 resentation can be recovered from the matrix-vector product $\mathbf{W}[1, 1, \dots, 1]^\top$, which sums
 1606 each row over the tokens $m = \{1, 2, \dots, M\}$. The matrix product $\Theta^{(x \rightarrow z)} \mathbf{W}$ contains the
 1607 horizontally concatenated embeddings of each word in the document, which will be use-
 1608 ful as the starting point for **convolutional neural networks** (see § 3.4). This is sometimes
 1609 called a **lookup layer**, because the first step is to lookup the embeddings for each word in
 1610 the input text.

1611 3.3 Learning neural networks

The feedforward network in Figure 3.1 can now be written as,

$$z \leftarrow f(\Theta^{(x \rightarrow z)} \mathbf{x}^{(i)}) \quad [3.21]$$

$$\tilde{y} \leftarrow \text{SoftMax}(\Theta^{(z \rightarrow y)} z + b) \quad [3.22]$$

$$\ell^{(i)} \leftarrow -e_{y^{(i)}} \cdot \log \tilde{y}, \quad [3.23]$$

where f is an elementwise activation function, such as σ or ReLU, and $\ell^{(i)}$ is the loss at instance i . The parameters $\Theta^{(x \rightarrow z)}$, $\Theta^{(z \rightarrow y)}$, and b can be estimated using online gradient-based optimization. The simplest such algorithm is stochastic gradient descent, which was discussed in § 2.6. Each parameter is updated by the gradient of the loss,

$$\mathbf{b} \leftarrow \mathbf{b} - \eta^{(t)} \nabla_{\mathbf{b}} \ell^{(i)} \quad [3.24]$$

$$\boldsymbol{\theta}_k^{(z \rightarrow y)} \leftarrow \boldsymbol{\theta}_k^{(z \rightarrow y)} - \eta^{(t)} \nabla_{\boldsymbol{\theta}_k^{(z \rightarrow y)}} \ell^{(i)} \quad [3.25]$$

$$\boldsymbol{\theta}_n^{(x \rightarrow z)} \leftarrow \boldsymbol{\theta}_n^{(x \rightarrow z)} - \eta^{(t)} \nabla_{\boldsymbol{\theta}_n^{(x \rightarrow z)}} \ell^{(i)}, \quad [3.26]$$

1612 where $\eta^{(t)}$ is the learning rate on iteration t , $\ell^{(i)}$ is the loss on instance (or minibatch) i ,
1613 and $\boldsymbol{\theta}_n^{(x \rightarrow z)}$ is column n of the matrix $\Theta^{(x \rightarrow z)}$, and $\boldsymbol{\theta}_k^{(z \rightarrow y)}$ is column k of $\Theta^{(z \rightarrow y)}$.

The gradients of the negative log-likelihood on \mathbf{b} and $\boldsymbol{\theta}_k^{(z \rightarrow y)}$ are similar to the gradients in logistic regression. For $\boldsymbol{\theta}_k^{(z \rightarrow y)}$, the gradient is,

$$\nabla_{\boldsymbol{\theta}_k^{(z \rightarrow y)}} \ell^{(i)} = \left[\frac{\partial \ell^{(i)}}{\partial \theta_{k,1}^{(z \rightarrow y)}}, \frac{\partial \ell^{(i)}}{\partial \theta_{k,2}^{(z \rightarrow y)}}, \dots, \frac{\partial \ell^{(i)}}{\partial \theta_{k,K_y}^{(z \rightarrow y)}} \right]^\top \quad [3.27]$$

$$\frac{\partial \ell^{(i)}}{\partial \theta_{k,j}^{(z \rightarrow y)}} = - \frac{\partial}{\partial \theta_{k,j}^{(z \rightarrow y)}} \left(\boldsymbol{\theta}_{y^{(i)}}^{(z \rightarrow y)} \cdot \mathbf{z} - \log \sum_{y \in \mathcal{Y}} \exp \boldsymbol{\theta}_y^{(z \rightarrow y)} \cdot \mathbf{z} \right) \quad [3.28]$$

$$= \left(\Pr(y = j \mid \mathbf{z}; \Theta^{(z \rightarrow y)}, \mathbf{b}) - \delta(j = y^{(i)}) \right) z_k, \quad [3.29]$$

1614 where $\delta(j = y^{(i)})$ is a function that returns one when $j = y^{(i)}$, and zero otherwise. The
1615 gradient $\nabla_{\mathbf{b}} \ell^{(i)}$ is similar to Equation 3.29.

The gradients on the input layer weights $\Theta^{(x \rightarrow z)}$ are obtained by the chain rule of differentiation:

$$\frac{\partial \ell^{(i)}}{\partial \theta_{n,k}^{(x \rightarrow z)}} = \frac{\partial \ell^{(i)}}{\partial z_k} \frac{\partial z_k}{\partial \theta_{n,k}^{(x \rightarrow z)}} \quad [3.30]$$

$$= \frac{\partial \ell^{(i)}}{\partial z_k} \frac{\partial f(\boldsymbol{\theta}_k^{(x \rightarrow z)} \cdot \mathbf{x})}{\partial \theta_{n,k}^{(x \rightarrow z)}} \quad [3.31]$$

$$= \frac{\partial \ell^{(i)}}{\partial z_k} \times f'(\boldsymbol{\theta}_k^{(x \rightarrow z)} \cdot \mathbf{x}) \times x_n, \quad [3.32]$$

where $f'(\boldsymbol{\theta}_k^{(x \rightarrow z)} \cdot \mathbf{x})$ is the derivative of the activation function f , applied at the input

$\theta_k^{(x \rightarrow z)} \cdot \mathbf{x}$. For example, if f is the sigmoid function, then the derivative is,

$$\frac{\partial \ell^{(i)}}{\partial \theta_{n,k}^{(x \rightarrow z)}} = \frac{\partial \ell^{(i)}}{\partial z_k} \times \sigma(\theta_k^{(x \rightarrow z)} \cdot \mathbf{x}) \times (1 - \sigma(\theta_k^{(x \rightarrow z)} \cdot \mathbf{x})) \times x_n \quad [3.33]$$

$$= \frac{\partial \ell^{(i)}}{\partial z_k} \times z_k \times (1 - z_k) \times x_n. \quad [3.34]$$

1616 For intuition, consider each of the terms in the product.

- 1617 • If the negative log-likelihood $\ell^{(i)}$ does not depend much on z_k , then $\frac{\partial \ell^{(i)}}{\partial z_k} \approx 0$. In this
1618 case it doesn't matter how z_k is computed, and so $\frac{\partial \ell^{(i)}}{\partial \theta_{n,k}^{(x \rightarrow z)}} \approx 0$.
- 1619 • If z_k is near 1 or 0, then the curve of the sigmoid function is nearly flat (Figure 3.2),
1620 and changing the inputs will make little local difference. The term $z_k \times (1 - z_k)$ is
1621 maximized at $z_k = \frac{1}{2}$, where the slope of the sigmoid function is steepest.
- 1622 • If $x_n = 0$, then it does not matter how we set the weights $\theta_{n,k}^{(x \rightarrow z)}$, so $\frac{\partial \ell^{(i)}}{\partial \theta_{n,k}^{(x \rightarrow z)}} = 0$.

1623 3.3.1 Backpropagation

1624 The equations above rely on the chain rule to compute derivatives of the loss with respect
1625 to each parameter of the model. Furthermore, local derivatives are frequently reused: for
1626 example, $\frac{\partial \ell^{(i)}}{\partial z_k}$ is reused in computing the derivatives with respect to each $\theta_{n,k}^{(x \rightarrow z)}$. These
1627 terms should therefore be computed once, and then cached. Furthermore, we should only
1628 compute any derivative once we have already computed all of the necessary “inputs”
1629 demanded by the chain rule of differentiation. This combination of sequencing, caching,
1630 and differentiation is known as **backpropagation**. It can be generalized to any directed
1631 acyclic **computation graph**.

1632 A computation graph is a declarative representation of a computational process. At
1633 each node t , compute a value v_t by applying a function f_t to a (possibly empty) list of
1634 parent nodes, π_t . Figure 3.3 shows the computation graph for a feedforward network
1635 with one hidden layer. There are nodes for the input $\mathbf{x}^{(i)}$, the hidden layer \mathbf{z} , the predicted
1636 output $\hat{\mathbf{y}}$, and the parameters Θ . During training, there is also a node for the ground truth
1637 label $y^{(i)}$ and the loss $\ell^{(i)}$. The predicted output $\hat{\mathbf{y}}$ is one of the parents of the loss (the
1638 other is the label $y^{(i)}$); its parents include Θ and \mathbf{z} , and so on.

1639 Computation graphs include three types of nodes:

1640 **Variables.** In the feedforward network of Figure 3.3, the variables include the inputs \mathbf{x} ,
1641 the hidden nodes \mathbf{z} , the outputs \mathbf{y} , and the loss function. Inputs are variables that
1642 do not have parents. Backpropagation computes the gradients with respect to all

Algorithm 6 General backpropagation algorithm. In the computation graph G , every node contains a function f_t and a set of parent nodes π_t ; the inputs to the graph are $x^{(i)}$.

variables except the inputs, and propagates these gradients backwards to the parameters.

1645 **Parameters.** In a feedforward network, the parameters include the weights and offsets.
 1646 In Figure 3.3, the parameters are summarized in the node Θ , but we could have
 1647 separate nodes for $\Theta^{(x \rightarrow z)}$, $\Theta^{(z \rightarrow y)}$, and any offset parameters. Parameter nodes do
 1648 not have parents; they are not computed from other nodes, but rather, are learned
 1649 by gradient descent.

Loss. The loss $\ell^{(i)}$ is the quantity that is to be minimized during training. The node representing the loss in the computation graph is not the parent of any other node; its parents are typically the predicted label \hat{y} and the true label $y^{(i)}$. Backpropagation begins by computing the gradient of the loss, and then propagating this gradient backwards to its immediate parents.

If the computation graph is a directed acyclic graph, then it is possible to order the nodes with a topological sort, so that if node t is a parent of node t' , then $t < t'$. This means that the values $\{v_t\}_{t=1}^T$ can be computed in a single forward pass. The topological sort is reversed when computing gradients: each gradient g_t is computed from the gradients of the children of t , implementing the chain rule of differentiation. The general backpropagation algorithm for computation graphs is shown in Algorithm 6.

While the gradients with respect to each parameter may be complex, they are composed of products of simple parts. For many networks, all gradients can be computed through **automatic differentiation**. This means that you need only specify the feedforward computation, and the gradients necessary for learning can be obtained automatically. There are many software libraries that perform automatic differentiation on compu-

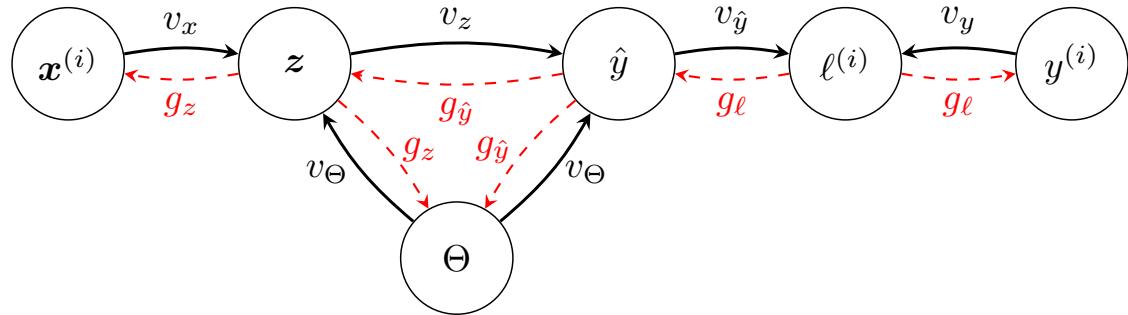


Figure 3.3: A computation graph for the feedforward neural network shown in Figure 3.1.

1666 tation graphs, such as TORCH (Collobert et al., 2011), TENSORFLOW (Abadi et al., 2016),
 1667 and DYNET (Neubig et al., 2017). One important distinction between these libraries is
 1668 whether they support **dynamic computation graphs**, in which the structure of the computa-
 1669 tion graph varies across instances. Static computation graphs are compiled in advance,
 1670 and can be applied to fixed-dimensional data, such as bag-of-words vectors. In many nat-
 1671 ural language processing problems, each input has a distinct structure, requiring a unique
 1672 computation graph. A simple case occurs in recurrent neural network language models
 1673 (see chapter 6), in which there is one node for each word in a sentence. More complex
 1674 cases include recursive neural networks (see chapter 14), in which the network is a tree
 1675 structure matching the syntactic organization of the input.

1676 3.3.2 Regularization and dropout

1677 In linear classification, overfitting was addressed by augmenting the objective with a reg-
 1678 ularization term, $\lambda \|\boldsymbol{\theta}\|_2^2$. This same approach can be applied to feedforward neural net-
 1679 works, penalizing each matrix of weights:

$$1680 L = \sum_{i=1}^N \ell^{(i)} + \lambda_{z \rightarrow y} \|\boldsymbol{\Theta}^{(z \rightarrow y)}\|_F^2 + \lambda_{x \rightarrow z} \|\boldsymbol{\Theta}^{(x \rightarrow z)}\|_F^2, \quad [3.35]$$

1681 where $\|\boldsymbol{\Theta}\|_F^2 = \sum_{i,j} \theta_{i,j}^2$ is the squared **Frobenius norm**, which generalizes the L_2 norm
 1682 to matrices. The bias parameters b are not regularized, as they do not contribute to the
 1683 sensitivity of the classifier to the inputs. In gradient-based optimization, the practical
 1684 effect of Frobenius norm regularization is that the weights “decay” towards zero at each
 update, motivating the alternative name **weight decay**.

1685 Another approach to controlling model complexity is **dropout**, which involves ran-
 1686 domly setting some computation nodes to zero during training (Srivastava et al., 2014).
 1687 For example, in the feedforward network, on each training instance, with probability ρ we

1688 set each input x_n and each hidden layer node z_k to zero. Srivastava et al. (2014) recom-
 1689 mend $\rho = 0.5$ for hidden units, and $\rho = 0.2$ for input units. Dropout is also incorporated
 1690 in the gradient computation, so if node z_k is dropped, then none of the weights $\theta_k^{(x \rightarrow z)}$ will
 1691 be updated for this instance. Dropout prevents the network from learning to depend too
 1692 much on any one feature or hidden node, and prevents **feature co-adaptation**, in which a
 1693 hidden unit is only useful in combination with one or more other hidden units. Dropout is
 1694 a special case of **feature noising**, which can also involve adding Gaussian noise to inputs
 1695 or hidden units (Holmstrom and Koistinen, 1992). Wager et al. (2013) show that dropout is
 1696 approximately equivalent to “adaptive” L_2 regularization, with a separate regularization
 1697 penalty for each feature.

1698 3.3.3 *Learning theory

1699 Chapter 2 emphasized the importance of **convexity** for learning: for convex objectives,
 1700 the global optimum can be found efficiently. The negative log-likelihood and hinge loss
 1701 are convex functions of the parameters of the output layer. However, the output of a feed-
 1702 forward network is generally not a convex function of the parameters of the input layer,
 1703 $\Theta^{(x \rightarrow z)}$. Feedforward networks can be viewed as function composition, where each layer
 1704 is a function that is applied to the output of the previous layer. Convexity is generally not
 1705 preserved in the composition of two convex functions — and furthermore, “squashing”
 1706 activation functions like tanh and sigmoid are not convex.

1707 The non-convexity of hidden layer neural networks can also be seen by permuting the
 1708 elements of the hidden layer, from $z = [z_1, z_2, \dots, z_{K_z}]$ to $\tilde{z} = [z_{\pi(1)}, z_{\pi(2)}, \dots, z_{\pi(K_z)}]$. This
 1709 corresponds to applying π to the rows of $\Theta^{(x \rightarrow z)}$ and the columns of $\Theta^{(z \rightarrow y)}$, resulting in
 1710 permuted parameter matrices $\Theta_\pi^{(x \rightarrow z)}$ and $\Theta_\pi^{(z \rightarrow y)}$. As long as this permutation is applied
 1711 consistently, the loss will be identical, $L(\Theta) = L(\Theta_\pi)$: it is *invariant* to this permutation.
 1712 However, the loss of the linear combination $L(\alpha\Theta + (1 - \alpha)\Theta_\pi)$ will generally not be
 1713 identical to the loss under Θ or its permutations. If $L(\Theta)$ is better than the loss at any
 1714 points in the immediate vicinity, and if $L(\Theta) = L(\Theta_\pi)$, then the loss function does not
 1715 satisfy the definition of convexity (see § 2.4). One of the exercises asks you to prove this
 1716 more rigorously.

1717 In practice, the existence of multiple optima is not necessarily problematic, if all such
 1718 optima are permutations of the sort described in the previous paragraph. In contrast,
 1719 “bad” local optima are better than their neighbors, but much worse than the global op-
 1720 timum. Fortunately, in large feedforward neural networks, most local optima are nearly
 1721 as good as the global optimum (Choromanska et al., 2015). More generally, a **critical**
 1722 **point** is one at which the gradient is zero. Critical points may be local optima, but they
 1723 may also be **saddle points**, which are local minima in some directions, but local *maxima*
 1724 in other directions. For example, the equation $x_1^2 - x_2^2$ has a saddle point at $x = (0, 0)$.
 1725 In large networks, the overwhelming majority of critical points are saddle points, rather

than local minima or maxima (Dauphin et al., 2014). Saddle points can pose problems for gradient-based optimization, since learning will slow to a crawl as the gradient goes to zero. However, the noise introduced by stochastic gradient descent, and by feature noising techniques such as dropout, can help online optimization to escape saddle points and find high-quality optima (Ge et al., 2015). Other techniques address saddle points directly, using local reconstructions of the Hessian matrix (Dauphin et al., 2014) or higher-order derivatives (Anandkumar and Ge, 2016).

Another theoretical puzzle about neural networks is how they are able to **generalize** to unseen data. Given enough parameters, a two-layer feedforward network can “memorize” its training data, attaining perfect accuracy on any training set. A particularly salient demonstration was provided by Zhang et al. (2017), who showed that neural networks can learn to perfectly classify a training set of images, even when the labels are replaced with random values! Of course, this network attains only chance accuracy when applied to heldout data. The concern is that when such a powerful learner is applied to real training data, it may learn a pathological classification function, which exploits irrelevant details of the training data and fails to generalize. Yet this extreme **overfitting** is rarely encountered in practice, and can usually be prevented by regularization, dropout, and early stopping (see § 3.3.4). Recent papers have derived generalization guarantees for specific classes of neural networks (e.g., Kawaguchi et al., 2017; Brutzkus et al., 2018), but theoretical work in this area is ongoing.

3.3.4 Tricks

Getting neural networks to work sometimes requires heuristic “tricks” (Bottou, 2012; Goodfellow et al., 2016; Goldberg, 2017b). This section presents some tricks that are especially important.

Initialization Initialization is not especially important for linear classifiers, since convexity ensures that the global optimum can usually be found quickly. But for multilayer neural networks, it is helpful to have a good starting point. One reason is that if the magnitude of the initial weights is too large, a sigmoid or tanh nonlinearity will be saturated, leading to a small gradient, and slow learning. Large gradients can cause training to diverge, with the parameters taking increasingly extreme values until reaching the limits of the floating point representation.

Initialization can help avoid these problems by ensuring that the variance over the initial gradients is constant and bounded throughout the network. For networks with tanh activation functions, this can be achieved by sampling the initial weights from the

following uniform distribution (Glorot and Bengio, 2010),

$$\theta_{i,j} \sim U \left[-\frac{\sqrt{6}}{\sqrt{d_{\text{in}}(n) + d_{\text{out}}(n)}}, \frac{\sqrt{6}}{\sqrt{d_{\text{in}}(n) + d_{\text{out}}(n)}} \right], \quad [3.36]$$

[3.37]

- 1757 For the weights leading to a ReLU activation function, He et al. (2015) use similar argu-
 1758 mentation to justify sampling from a zero-mean Gaussian distribution,

$$\theta_{i,j} \sim N(0, \sqrt{2/d_{\text{in}}(n)}) \quad [3.38]$$

Rather than initializing the weights independently, it can be beneficial to initialize each layer jointly as an **orthonormal matrix**, ensuring that $\Theta^\top \Theta = \mathbb{I}$ (Saxe et al., 2014). Orthonormal matrices preserve the norm of the input, so that $\|\Theta x\| = \|x\|$, which prevents the gradients from exploding or vanishing. Orthogonality ensures that the hidden units are uncorrelated, so that they correspond to different features of the input. Orthonormal initialization can be performed by applying **singular value decomposition** to a matrix of values sampled from a standard normal distribution:

$$a_{i,j} \sim N(0, 1) \quad [3.39]$$

$$\mathbf{A} = \{a_{i,j}\}_{i=1,j=1}^{d_{\text{in}}(j), d_{\text{out}}(j)} \quad [3.40]$$

$$\mathbf{U}, \mathbf{S}, \mathbf{V}^\top = \text{SVD}(\mathbf{A}) \quad [3.41]$$

$$\Theta^{(j)} \leftarrow \mathbf{U}. \quad [3.42]$$

- 1759 The matrix \mathbf{U} contains the **singular vectors** of \mathbf{A} , and is guaranteed to be orthonormal.
 1760 For more on singular value decomposition, see chapter 14.

- 1761 Even with careful initialization, there can still be significant variance in the final re-
 1762 sults. It can be useful to make multiple training runs, and select the one with the best
 1763 performance on a heldout development set.

- 1764 **Clipping and normalization** Learning can be sensitive to the magnitude of the gradient:
 1765 too large, and learning can diverge, with successive updates thrashing between increas-
 1766 ingly extreme values; too small, and learning can grind to a halt. Several heuristics have
 1767 been proposed to address this issue.

- 1768 • In **gradient clipping** (Pascanu et al., 2013), an upper limit is placed on the norm of
 1769 the gradient, and the gradient is rescaled when this limit is exceeded,

$$\text{CLIP}(\hat{\mathbf{g}}) = \begin{cases} \mathbf{g} & \|\hat{\mathbf{g}}\| < \tau \\ \frac{\tau}{\|\mathbf{g}\|} \mathbf{g} & \text{otherwise.} \end{cases} \quad [3.43]$$

- In **batch normalization** (Ioffe and Szegedy, 2015), the inputs to each computation node are recentered by their mean and variance across all of the instances in the minibatch \mathcal{B} (see § 2.6.2). For example, in a feedforward network with one hidden layer, batch normalization would transform the inputs to the hidden layer as follows:

$$\boldsymbol{\mu}^{(\mathcal{B})} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \mathbf{x}^{(i)} \quad [3.44]$$

$$\mathbf{s}^{(\mathcal{B})} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} (\mathbf{x}^{(i)} - \boldsymbol{\mu}^{(\mathcal{B})})^2 \quad [3.45]$$

$$\bar{\mathbf{x}}^{(i)} = (\mathbf{x}^{(i)} - \boldsymbol{\mu}^{(\mathcal{B})}) / \sqrt{\mathbf{s}^{(\mathcal{B})}}. \quad [3.46]$$

Empirically, this speeds convergence of deep architectures. One explanation is that it helps to correct for changes in the distribution of activations during training.

- In **layer normalization** (Ba et al., 2016), the inputs to each nonlinear activation function are recentered across the layer:

$$\mathbf{a} = \Theta^{(x \rightarrow z)} \mathbf{x} \quad [3.47]$$

$$\mu = \frac{1}{K_z} \sum_{k=1}^{K_z} a_k \quad [3.48]$$

$$s = \frac{1}{K_z} \sum_{k=1}^{K_z} (a_k - \mu)^2 \quad [3.49]$$

$$\mathbf{z} = (\mathbf{a} - \mu) / \sqrt{s}. \quad [3.50]$$

Layer normalization has similar motivations to batch normalization, but it can be applied across a wider range of architectures and training conditions.

Online optimization There is a cottage industry of online optimization algorithms that attempt to improve on stochastic gradient descent. **AdaGrad** was reviewed in § 2.6.2; its main innovation is to set adaptive learning rates for each parameter by storing the sum of squared gradients. Rather than using the sum over the entire training history, we can keep a running estimate,

$$v_j^{(t)} = \beta v_j^{(t-1)} + (1 - \beta) g_{t,j}^2, \quad [3.51]$$

where $g_{t,j}$ is the gradient with respect to parameter j at time t , and $\beta \in [0, 1]$. This term places more emphasis on recent gradients, and is employed in the AdaDelta (Zeiler, 2012) and Adam (Kingma and Ba, 2014) optimizers. Online optimization and its theoretical background are reviewed by Bottou et al. (2016). **Early stopping**, mentioned in § 2.3.2, can help to avoid overfitting by terminating training after reaching a plateau in the performance on a heldout validation set.

1780 **Practical advice** The bag of tricks for training neural networks continues to grow, and
 1781 it is likely that there will be several new ones by the time you read this. Today, it is
 1782 standard practice to use gradient clipping, early stopping, and a sensible initialization of
 1783 parameters to small random values. More bells and whistles can be added as solutions to
 1784 specific problems — for example, if it is difficult to find a good learning rate for stochastic
 1785 gradient descent, then it may help to try a fancier optimizer with an adaptive learning
 1786 rate. Alternatively, if a method such as layer normalization is used by related models
 1787 in the research literature, you should probably consider it, especially if you are having
 1788 trouble matching published results. As with linear classifiers, it is important to evaluate
 1789 these decisions on a held-out development set, and not on the test set that will be used to
 1790 provide the final measure of the model’s performance (see § 2.2.5).

1791 3.4 Convolutional neural networks

1792 A basic weakness of the bag-of-words model is its inability to account for the ways in
 1793 which words combine to create meaning, including even simple reversals such as *not*
pleasant, hardly a generous offer, and *I wouldn’t mind missing the flight*. Computer vision
 1794 faces the related challenge of identifying the semantics of images from pixel features
 1795 that are uninformative in isolation. An earlier generation of computer vision research
 1796 focused on designing *filters* to aggregate local pixel-level features into more meaningful
 1797 representations, such as edges and corners (e.g., Canny, 1987). Similarly, earlier NLP re-
 1798 search attempted to capture multiword linguistic phenomena by hand-designed lexical
 1799 patterns (Hobbs et al., 1997). In both cases, the output of the filters and patterns could
 1800 then act as base features in a linear classifier. But rather than designing these feature ex-
 1801 tractors by hand, a better approach is to learn them, using the magic of backpropagation.
 1802 This is the idea behind **convolutional neural networks**.

1803 Following § 3.2.4, define the base layer of a neural network as,

$$\mathbf{X}^{(0)} = \Theta^{(x \rightarrow z)}[e_{w_1}, e_{w_2}, \dots, e_{w_M}], \quad [3.52]$$

1804 where e_{w_m} is a column vector of zeros, with a 1 at position w_m . The base layer has dimension
 $\mathbf{X}^{(0)} \in \mathbb{R}^{K_e \times M}$, where K_e is the size of the word embeddings. To merge information
 across adjacent words, we *convolve* $\mathbf{X}^{(0)}$ with a set of filter matrices $\mathbf{C}^{(k)} \in \mathbb{R}^{K_e \times h}$. Convo-
 lution is indicated by the symbol $*$, and is defined,

$$\mathbf{X}^{(1)} = f(\mathbf{b} + \mathbf{C} * \mathbf{X}^{(0)}) \implies x_{k,m}^{(1)} = f \left(b_k + \sum_{k'=1}^{K_e} \sum_{n=1}^h c_{k',n}^{(k)} \times x_{k',m+n-1}^{(0)} \right), \quad [3.53]$$

1805 where f is an activation function such as tanh or ReLU, and \mathbf{b} is a vector of offsets. The
 1806 convolution operation slides the matrix $\mathbf{C}^{(k)}$ across the columns of $\mathbf{X}^{(0)}$. At each position
 1807 m , we compute the elementwise product $\mathbf{C}^{(k)} \odot \mathbf{X}_{m:m+h-1}^{(0)}$, and take the sum.

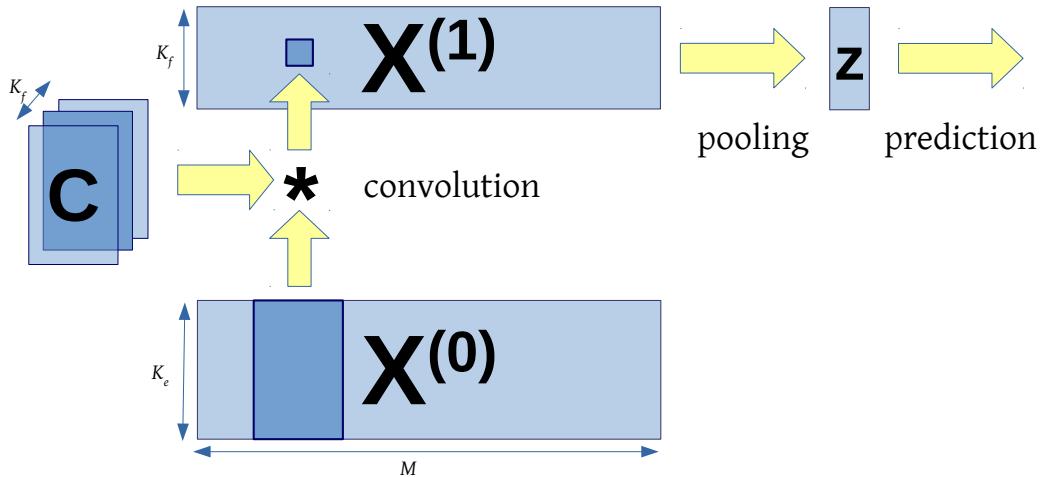


Figure 3.4: A convolutional neural network for text classification

1808 A simple filter might compute a weighted average over nearby words,

$$C^{(k)} = \begin{bmatrix} 0.5 & 1 & 0.5 \\ 0.5 & 1 & 0.5 \\ \dots & \dots & \dots \\ 0.5 & 1 & 0.5 \end{bmatrix}, \quad [3.54]$$

1809 thereby representing trigram units like *not so unpleasant*. In **one-dimensional convolution**,
 1810 each filter matrix $C^{(k)}$ is constrained to have non-zero values only at row k (Kalch-
 1811 brenner et al., 2014). This means that each dimension of the word embedding is processed
 1812 by a separate filter, and it implies that $K_f = K_e$.

1813 To deal with the beginning and end of the input, the base matrix $X^{(0)}$ may be padded
 1814 with h column vectors of zeros at the beginning and end; this is known as **wide convolution**.
 1815 If padding is not applied, then the output from each layer will be $h - 1$ units smaller
 1816 than the input; this is known as **narrow convolution**. The filter matrices need not have
 1817 identical filter widths, so more generally we could write h_k to indicate the width of filter
 1818 $C^{(k)}$. As suggested by the notation $X^{(0)}$, multiple layers of convolution may be applied,
 1819 so that $X^{(d)}$ is the input to $X^{(d+1)}$.

After D convolutional layers, we obtain a matrix representation of the document $X^{(D)} \in \mathbb{R}^{K_z \times M}$. If the instances have variable lengths, it is necessary to aggregate over all M word positions to obtain a fixed-length representation. This can be done by a **pooling** operation,

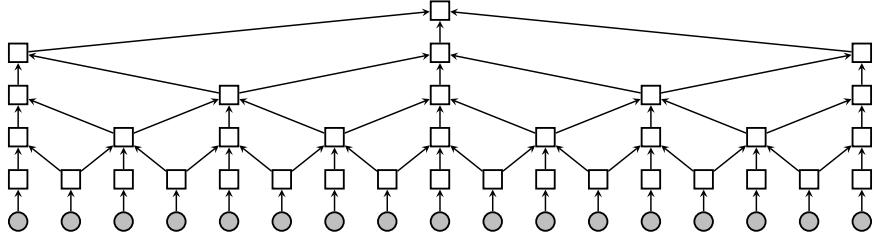


Figure 3.5: A dilated convolutional neural network captures progressively larger context through recursive application of the convolutional operator

such as max-pooling (Collobert et al., 2011) or average-pooling,

$$\mathbf{z} = \text{MaxPool}(\mathbf{X}^{(D)}) \implies z_k = \max(x_{k,1}^{(D)}, x_{k,2}^{(D)}, \dots, x_{k,M}^{(D)}) \quad [3.55]$$

$$\mathbf{z} = \text{AvgPool}(\mathbf{X}^{(D)}) \implies z_k = \frac{1}{M} \sum_{m=1}^M x_{k,m}^{(D)}. \quad [3.56]$$

1820 The vector \mathbf{z} can now act as a layer in a feedforward network, culminating in a prediction
1821 \hat{y} and a loss $\ell^{(i)}$. The setup is shown in Figure 3.4.

Just as in feedforward networks, the parameters $(\mathbf{C}^{(k)}, \mathbf{b}, \Theta)$ can be learned by backpropagating from the classification loss. This requires backpropagating through the max-pooling operation, which is a discontinuous function of the input. But because we need only a local gradient, backpropagation flows only through the argmax m :

$$\frac{\partial z_k}{\partial x_{k,m}^{(D)}} = \begin{cases} 1, & x_{k,m}^{(D)} = \max(x_{k,1}^{(D)}, x_{k,2}^{(D)}, \dots, x_{k,M}^{(D)}) \\ 0, & \text{otherwise.} \end{cases} \quad [3.57]$$

1822 The computer vision literature has produced a huge variety of convolutional archi-
1823 tectures, and many of these innovations can be applied to text data. One avenue for
1824 improvement is more complex pooling operations, such as k -max pooling (Kalchbrenner
1825 et al., 2014), which returns a matrix of the k largest values for each filter. Another innova-
1826 tion is the use of **dilated convolution** to build multiscale representations (Yu and Koltun,
1827 2016). At each layer, the convolutional operator applied in *strides*, skipping ahead by s
1828 steps after each feature. As we move up the hierarchy, each layer is s times smaller than
1829 the layer below it, effectively summarizing the input (Kalchbrenner et al., 2016; Strubell
1830 et al., 2017). This idea is shown in Figure 3.5. Multi-layer convolutional networks can also
1831 be augmented with “shortcut” connections, as in the residual network from § 3.2.2 (John-
1832 son and Zhang, 2017).

1833 **Additional resources**

1834 The deep learning textbook by Goodfellow et al. (2016) covers many of the topics in this
 1835 chapter in more detail. For a comprehensive review of neural networks in natural lan-
 1836 guage processing, see Goldberg (2017b). A seminal work on deep learning in natural
 1837 language processing is the aggressively titled “Natural Language Processing (Almost)
 1838 from Scratch”, which uses convolutional neural networks to perform a range of language
 1839 processing tasks (Collobert et al., 2011), although there is earlier work (e.g., Henderson,
 1840 2004). This chapter focuses on feedforward and convolutional neural networks, but recur-
 1841 rent neural networks are one of the most important deep learning architectures for natural
 1842 language processing. They are covered extensively in chapters 6 and 7.

1843 The role of deep learning in natural language processing research has caused angst
 1844 in some parts of the natural language processing research community (e.g., Goldberg,
 1845 2017a), especially as some of the more zealous deep learning advocates have argued that
 1846 end-to-end learning from “raw” text can eliminate the need for linguistic constructs such
 1847 as sentences, phrases, and even words (Zhang et al., 2015, originally titled “Text under-
 1848 standing from scratch”). These developments were surveyed by Manning (2015). While
 1849 reports of the demise of linguistics in natural language processing remain controversial
 1850 at best, deep learning and backpropagation have become ubiquitous in both research and
 1851 applications.

1852 **Exercises**

- 1853 1. Figure 3.3 shows the computation graph for a feedforward neural network with one
 1854 layer.
 - 1855 a) Update the computation graph to include a residual connection between x and
 1856 z .
 - 1857 b) Update the computation graph to include a highway connection between x
 1858 and z .
- 1859 2. Prove that the softmax and sigmoid functions are equivalent when the number of
 1860 possible labels is two. Specifically, for any $\Theta^{(z \rightarrow y)}$ (omitting the offset b for simplic-
 1861 ity), show how to construct a vector of weights θ such that,

$$\text{SoftMax}(\Theta^{(z \rightarrow y)} z)[0] = \sigma(\theta \cdot z). \quad [3.58]$$

- 1862 3. Convolutional neural networks often aggregate across words by using **max-pooling**
 1863 (Equation 3.55 in § 3.4). A potential concern is that there is zero gradient with re-
 1864 spect to the parts of the input that are not included in the maximum. The following

questions consider the gradient with respect to an element of the input, $x_{m,k}^{(0)}$, and they assume that all parameters are independently distributed.

- a) First consider a minimal network, with $z = \text{MaxPool}(\mathbf{X}^{(0)})$. What is the probability that the gradient $\frac{\partial \ell}{\partial x_{m,k}^{(0)}}$ is non-zero?
- b) Now consider a two-level network, with $\mathbf{X}^{(1)} = f(\mathbf{b} + \mathbf{C} * \mathbf{X}^{(0)})$. Express the probability that the gradient $\frac{\partial \ell}{\partial x_{m,k}^{(0)}}$ is non-zero, in terms of the input length M , the filter size n , and the number of filters K_f .
- c) Using a calculator, work out the probability for the case $M = 128, n = 4, K_f = 32$.
- d) Now consider a three-level network, $\mathbf{X}^{(2)} = f(\mathbf{b} + \mathbf{C} * \mathbf{X}^{(1)})$. Give the general equation for the probability that $\frac{\partial \ell}{\partial x_{m,k}^{(0)}}$ is non-zero, and compute the numerical probability for the scenario in the previous part, assuming $K_f = 32$ and $n = 4$ at both levels.

4. Design a feedforward network to compute the XOR function:

$$f(x_1, x_2) = \begin{cases} -1, & x_1 = 1, x_2 = 1 \\ 1, & x_1 = 1, x_2 = 0 \\ 1, & x_1 = 0, x_2 = 1 \\ -1, & x_1 = 0, x_2 = 0 \end{cases}. \quad [3.59]$$

Your network should have a single output node which uses the Sign activation function, $f(x) = \begin{cases} 1, & x > 0 \\ -1, & x \leq 0. \end{cases}$. Use a single hidden layer, with ReLU activation functions. Describe all weights and offsets.

- 5. Consider the same network as above (with ReLU activations for the hidden layer), with an arbitrary differentiable loss function $\ell(y^{(i)}, \tilde{y})$, where \tilde{y} is the activation of the output node. Suppose all weights and offsets are initialized to zero. Show that gradient descent will not learn the desired function from this initialization.
- 6. The simplest solution to the previous problem relies on the use of the ReLU activation function at the hidden layer. Now consider a network with arbitrary activations on the hidden layer. Show that if the initial weights are any uniform constant, then gradient descent will not learn the desired function from this initialization.
- 7. Consider a network in which: the base features are all binary, $\mathbf{x} \in \{0, 1\}^M$; the hidden layer activation function is sigmoid, $z_k = \sigma(\theta_k \cdot \mathbf{x})$; and the initial weights are sampled independently from a standard normal distribution, $\theta_{j,k} \sim N(0, 1)$.

- 1893 • Show how the probability of a small initial gradient on any weight, $\frac{\partial z_k}{\partial \theta_{j,k}} < \alpha$,
 1894 depends on the size of the input M . Hint: use the lower bound,

$$\Pr(\sigma(\theta_k \cdot \mathbf{x}) \times (1 - \sigma(\theta_k \cdot \mathbf{x})) < \alpha) \geq 2 \Pr(\sigma(\theta_k \cdot \mathbf{x}) < \alpha), \quad [3.60]$$

1895 and relate this probability to the variance $V[\theta_k \cdot \mathbf{x}]$.

- 1896 • Design an alternative initialization that removes this dependence.

8. The ReLU activation function can lead to “dead neurons”, which can never be activated on any input. Consider the following two-layer feedforward network with a scalar output y :

$$z_i = \text{ReLU}(\theta_i^{(x \rightarrow z)} \cdot \mathbf{x} + b_i) \quad [3.61]$$

$$y = \theta^{(z \rightarrow y)} \cdot \mathbf{z}. \quad [3.62]$$

1897 Suppose that the input is a binary vector of observations, $\mathbf{x} \in \{0, 1\}^D$.

- 1898 a) Under what condition is node z_i “dead”? Your answer should be expressed in
 1899 terms of the parameters $\theta_i^{(x \rightarrow z)}$ and b_i .

- 1900 b) Suppose that the gradient of the loss on a given instance is $\frac{\partial \ell}{\partial y} = 1$. Derive the
 1901 gradients $\frac{\partial \ell}{\partial b_i}$ and $\frac{\partial \ell}{\partial \theta_{j,i}^{(x \rightarrow z)}}$ for such an instance.

- 1902 c) Using your answers to the previous two parts, explain why a dead neuron can
 1903 never be brought back to life during gradient-based learning.

9. Suppose that the parameters $\Theta = \{\Theta^{(x \rightarrow z)}, \Theta(z \rightarrow y), \mathbf{b}\}$ are a local optimum of a feedforward network in the following sense: there exists some $\epsilon > 0$ such that,

$$\begin{aligned} & \left(\|\tilde{\Theta}^{(x \rightarrow z)} - \Theta^{(x \rightarrow z)}\|_F^2 + \|\tilde{\Theta}^{(z \rightarrow y)} - \Theta^{(z \rightarrow y)}\|_F^2 + \|\tilde{\mathbf{b}} - \mathbf{b}\|_2^2 < \epsilon \right) \\ & \Rightarrow \left(L(\tilde{\Theta}) > L(\Theta) \right) \end{aligned} \quad [3.63]$$

1904 Define the function π as a permutation on the hidden units, as described in § 3.3.3,
 1905 so that for any Θ , $L(\Theta) = L(\Theta_\pi)$. Prove that if a feedforward network has a local
 1906 optimum in the sense of Equation 3.63, then its loss is not a convex function of the
 1907 parameters Θ , using the definition of convexity from § 2.4

- 1908 10. Consider a network with a single hidden layer, and a single output,

$$y = \theta^{(z \rightarrow y)} \cdot g(\Theta^{(x \rightarrow z)} \mathbf{x}). \quad [3.64]$$

1909 Assume that g is the ReLU function. Show that for any matrix of weights $\Theta^{(x \rightarrow z)}$, it
 1910 is permissible to rescale each row to have a norm of one, because an identical output
 1911 can be obtained by finding a corresponding rescaling of $\theta^{(z \rightarrow y)}$.

¹⁹¹² Chapter 4

¹⁹¹³ **Linguistic applications of ¹⁹¹⁴ classification**

¹⁹¹⁵ Having covered several techniques for classification, this chapter shifts the focus from
¹⁹¹⁶ mathematics to linguistic applications. Later in the chapter, we will consider the design
¹⁹¹⁷ decisions involved in text classification, as well as best practices for evaluation.

¹⁹¹⁸ 4.1 Sentiment and opinion analysis

¹⁹¹⁹ A popular application of text classification is to automatically determine the **sentiment**
¹⁹²⁰ or **opinion polarity** of documents such as product reviews and social media posts. For
¹⁹²¹ example, marketers are interested to know how people respond to advertisements, ser-
¹⁹²² vices, and products (Hu and Liu, 2004); social scientists are interested in how emotions
¹⁹²³ are affected by phenomena such as the weather (Hannak et al., 2012), and how both opin-
¹⁹²⁴ ions and emotions spread over social networks (Coviello et al., 2014; Miller et al., 2011).
¹⁹²⁵ In the field of **digital humanities**, literary scholars track plot structures through the flow
¹⁹²⁶ of sentiment across a novel (Jockers, 2015).¹

¹⁹²⁷ Sentiment analysis can be framed as a direct application of document classification,
¹⁹²⁸ assuming reliable labels can be obtained. In the simplest case, sentiment analysis is a
¹⁹²⁹ two or three-class problem, with sentiments of POSITIVE, NEGATIVE, and possibly NEU-
¹⁹³⁰ TRAL. Such annotations could be annotated by hand, or obtained automatically through
¹⁹³¹ a variety of means:

- ¹⁹³² • Tweets containing happy emoticons can be marked as positive, sad emoticons as
¹⁹³³ negative (Read, 2005; Pak and Paroubek, 2010).

¹Comprehensive surveys on sentiment analysis and related problems are offered by Pang and Lee (2008) and Liu (2015).

- 1934 • Reviews with four or more stars can be marked as positive, three or fewer stars as
1935 negative (Pang et al., 2002).
- 1936 • Statements from politicians who are voting for a given bill are marked as positive
1937 (towards that bill); statements from politicians voting against the bill are marked as
1938 negative (Thomas et al., 2006).

1939 The bag-of-words model is a good fit for sentiment analysis at the document level: if
1940 the document is long enough, we would expect the words associated with its true senti-
1941 ment to overwhelm the others. Indeed, **lexicon-based sentiment analysis** avoids machine
1942 learning altogether, and classifies documents by counting words against positive and neg-
1943 ative sentiment word lists (Taboada et al., 2011).

1944 Lexicon-based classification is less effective for short documents, such as single-sentence
1945 reviews or social media posts. In these documents, linguistic issues like **negation** and **ir-**
1946 **realis** (Polanyi and Zaenen, 2006) — events that are hypothetical or otherwise non-factual
1947 — can make bag-of-words classification ineffective. Consider the following examples:

- 1948 (4.1) a. That's not bad for the first day.
1949 b. This is not the worst thing that can happen.
1950 c. It would be nice if you acted like you understood.
1951 d. There is no reason at all to believe that the polluters are suddenly going to
1952 become reasonable. (Wilson et al., 2005)
1953 e. This film should be brilliant. The actors are first grade. Stallone plays a
1954 happy, wonderful man. His sweet wife is beautiful and adores him. He has
1955 a fascinating gift for living life fully. It sounds like a great plot, **however**, the
1956 film is a failure. (Pang et al., 2002)

1957 A minimal solution is to move from a bag-of-words model to a bag-of-**bigrams** model,
1958 where each base feature is a pair of adjacent words, e.g.,

$$(that's, not), (not, bad), (bad, for), \dots \quad [4.1]$$

1959 Bigrams can handle relatively straightforward cases, such as when an adjective is imme-
1960 diately negated; trigrams would be required to extend to larger contexts (e.g., *not the worst*).
1961 But this approach will not scale to more complex examples like (4.1d) and (4.1e). More
1962 sophisticated solutions try to account for the syntactic structure of the sentence (Wilson
1963 et al., 2005; Socher et al., 2013), or apply more complex classifiers such as convolutional
1964 neural networks (Kim, 2014), which are described in chapter 3.

1965 4.1.1 Related problems

1966 **Subjectivity** Closely related to sentiment analysis is **subjectivity detection**, which re-
1967 quires identifying the parts of a text that express subjective opinions, as well as other non-

1968 factual content such as speculation and hypotheticals (Riloff and Wiebe, 2003). This can be
1969 done by treating each sentence as a separate document, and then applying a bag-of-words
1970 classifier: indeed, Pang and Lee (2004) do exactly this, using a training set consisting of
1971 (mostly) subjective sentences gathered from movie reviews, and (mostly) objective sen-
1972 tences gathered from plot descriptions. They augment this bag-of-words model with a
1973 graph-based algorithm that encourages nearby sentences to have the same subjectivity
1974 label.

1975 **Stance classification** In debates, each participant takes a side: for example, advocating
1976 for or against proposals like adopting a vegetarian lifestyle or mandating free college ed-
1977 ucation. The problem of stance classification is to identify the author’s position from the
1978 text of the argument. In some cases, there is training data available for each position,
1979 so that standard document classification techniques can be employed. In other cases, it
1980 suffices to classify each document as whether it is in support or opposition of the argu-
1981 ment advanced by a previous document (Anand et al., 2011). In the most challenging
1982 case, there is no labeled data for any of the stances, so the only possibility is group docu-
1983 ments that advocate the same position (Somasundaran and Wiebe, 2009). This is a form
1984 of **unsupervised learning**, discussed in chapter 5.

1985 **Targeted sentiment analysis** The expression of sentiment is often more nuanced than a
1986 simple binary label. Consider the following examples:

- 1987 (4.2) a. The vodka was good, but the meat was rotten.
1988 b. Go to Heaven for the climate, Hell for the company. —Mark Twain

1989 These statements display a mixed overall sentiment: positive towards some entities (e.g.,
1990 *the vodka*), negative towards others (e.g., *the meat*). **Targeted sentiment analysis** seeks to
1991 identify the writer’s sentiment towards specific entities (Jiang et al., 2011). This requires
1992 identifying the entities in the text and linking them to specific sentiment words — much
1993 more than we can do with the classification-based approaches discussed thus far. For
1994 example, Kim and Hovy (2006) analyze sentence-internal structure to determine the topic
1995 of each sentiment expression.

1996 **Aspect-based opinion mining** seeks to identify the sentiment of the author of a review
1997 towards predefined aspects such as PRICE and SERVICE, or, in the case of (4.2b), CLIMATE
1998 and COMPANY (Hu and Liu, 2004). If the aspects are not defined in advance, it may again
1999 be necessary to employ unsupervised learning methods to identify them (e.g., Branavan
2000 et al., 2009).

2001 **Emotion classification** While sentiment analysis is framed in terms of positive and neg-
2002 ative categories, psychologists generally regard **emotion** as more multifaceted. For ex-
2003 ample, Ekman (1992) argues that there are six basic emotions — happiness, surprise, fear,

2004 sadness, anger, and contempt — and that they are universal across human cultures. Alm
 2005 et al. (2005) build a linear classifier for recognizing the emotions expressed in children’s
 2006 stories. The ultimate goal of this work was to improve text-to-speech synthesis, so that
 2007 stories could be read with intonation that reflected the emotional content. They used bag-
 2008 of-words features, as well as features capturing the story type (e.g., jokes, folktales), and
 2009 structural features that reflect the position of each sentence in the story. The task is diffi-
 2010 cult: even human annotators frequently disagreed with each other, and the best classifiers
 2011 achieved accuracy between 60-70%.

2012 4.1.2 Alternative approaches to sentiment analysis

2013 **Regression** A more challenging version of sentiment analysis is to determine not just
 2014 the class of a document, but its rating on a numerical scale (Pang and Lee, 2005). If the
 2015 scale is continuous, it is most natural to apply **regression**, identifying a set of weights θ
 2016 that minimize the squared error of a predictor $\hat{y} = \theta \cdot x + b$, where b is an offset. This
 2017 approach is called **linear regression**, and sometimes **least squares**, because the regression
 2018 coefficients θ are determined by minimizing the squared error, $(y - \hat{y})^2$. If the weights are
 2019 regularized using a penalty $\lambda \|\theta\|_2^2$, then it is **ridge regression**. Unlike logistic regression,
 2020 both linear regression and ridge regression can be solved in closed form as a system of
 2021 linear equations.

2022 **Ordinal ranking** In many problems, the labels are ordered but discrete: for example,
 2023 product reviews are often integers on a scale of 1 – 5, and grades are on a scale of A – F.
 2024 Such problems can be solved by discretizing the score $\theta \cdot x$ into “ranks”,

$$\hat{y} = \underset{r: \theta \cdot x \geq b_r}{\operatorname{argmax}} r, \quad [4.2]$$

2025 where $\mathbf{b} = [b_1 = -\infty, b_2, b_3, \dots, b_K]$ is a vector of boundaries. It is possible to learn the
 2026 weights and boundaries simultaneously, using a perceptron-like algorithm (Crammer and
 2027 Singer, 2001).

2028 **Lexicon-based classification** Sentiment analysis is one of the only NLP tasks where
 2029 hand-crafted feature weights are still widely employed. In **lexicon-based classification** (Taboada
 2030 et al., 2011), the user creates a list of words for each label, and then classifies each docu-
 2031 ment based on how many of the words from each list are present. In our linear classifica-
 2032 tion framework, this is equivalent to choosing the following weights:

$$\theta_{y,j} = \begin{cases} 1, & j \in \mathcal{L}_y \\ 0, & \text{otherwise,} \end{cases} \quad [4.3]$$

2033 where \mathcal{L}_y is the lexicon for label y . Compared to the machine learning classifiers discussed
 2034 in the previous chapters, lexicon-based classification may seem primitive. However, su-
 2035 pervised machine learning relies on large annotated datasets, which are time-consuming
 2036 and expensive to produce. If the goal is to distinguish two or more categories in a new
 2037 domain, it may be simpler to start by writing down a list of words for each category.

2038 An early lexicon was the *General Inquirer* (Stone, 1966). Today, popular sentiment lex-
 2039 cons include SENTIWORDNET (Esuli and Sebastiani, 2006) and an evolving set of lexicons
 2040 from Liu (2015). For emotions and more fine-grained analysis, *Linguistic Inquiry and Word*
 2041 *Count* (LIWC) provides a set of lexicons (Tausczik and Pennebaker, 2010). The MPQA lex-
 2042 icon indicates the polarity (positive or negative) of 8221 terms, as well as whether they are
 2043 strongly or weakly subjective (Wiebe et al., 2005). A comprehensive comparison of senti-
 2044 ment lexicons is offered by Ribeiro et al. (2016). Given an initial **seed lexicon**, it is possible
 2045 to automatically expand the lexicon by looking for words that frequently co-occur with
 2046 words in the seed set (Hatzivassiloglou and McKeown, 1997; Qiu et al., 2011).

2047 4.2 Word sense disambiguation

2048 Consider the the following headlines:

- 2049 (4.3) a. Iraqi head seeks arms
 2050 b. Prostitutes appeal to Pope
 2051 c. Drunk gets nine years in violin case²

2052 These headlines are ambiguous because they contain words that have multiple mean-
 2053 ings, or **senses**. Word sense disambiguation is the problem of identifying the intended
 2054 sense of each word token in a document. Word sense disambiguation is part of a larger
 2055 field of research called **lexical semantics**, which is concerned with meanings of the words.

2056 At a basic level, the problem of word sense disambiguation is to identify the correct
 2057 sense for each word token in a document. Part-of-speech ambiguity (e.g., noun versus
 2058 verb) is usually considered to be a different problem, to be solved at an earlier stage.
 2059 From a linguistic perspective, senses are not properties of words, but of **lemmas**, which
 2060 are canonical forms that stand in for a set of inflected words. For example, *arm*/N is a
 2061 lemma that includes the inflected form *arms*/N — the /N indicates that it we are refer-
 2062 ring to the noun, and not its **homonym** *arm*/V, which is another lemma that includes
 2063 the inflected verbs (*arm*/V, *arms*/V, *armed*/V, *arming*/V). Therefore, word sense disam-
 2064 biguation requires first identifying the correct part-of-speech and lemma for each token,

²These examples, and many more, can be found at <http://www.ling.upenn.edu/~beatrice/humor/headlines.html>

2065 and then choosing the correct sense from the inventory associated with the corresponding
 2066 lemma.³ (Part-of-speech tagging is discussed in § 8.1.)

2067 **4.2.1 How many word senses?**

2068 Words sometimes have many more than two senses, as exemplified by the word *serve*:

- 2069 • [FUNCTION]: *The tree stump served as a table*
- 2070 • [CONTRIBUTE TO]: *His evasive replies only served to heighten suspicion*
- 2071 • [PROVIDE]: *We serve only the rawest fish*
- 2072 • [ENLIST]: *She served in an elite combat unit*
- 2073 • [JAIL]: *He served six years for a crime he didn't commit*
- 2074 • [LEGAL]: *They were served with subpoenas*⁴

2075 These sense distinctions are annotated in WORDNET (<http://wordnet.princeton.edu>).
 2076 WORDNET consists of roughly 100,000
 2077 **synsets**, which are groups of lemmas (or phrases) that are synonymous. An example
 2078 synset is {*chump*¹, *fool*², *sucker*¹, *mark*⁹}, where the superscripts index the sense of each
 2079 lemma that is included in the synset: for example, there are at least eight other senses of
 2080 *mark* that have different meanings, and are not part of this synset. A lemma is **polysemous**
 2081 if it participates in multiple synsets.

2082 WORDNET defines the scope of the word sense disambiguation problem, and, more
 2083 generally, formalizes lexical semantic knowledge of English. (WordNets have been cre-
 2084 ated for a few dozen other languages, at varying levels of detail.) Some have argued
 2085 that WordNet's sense granularity is too fine (Ide and Wilks, 2006); more fundamentally,
 2086 the premise that word senses can be differentiated in a task-neutral way has been criti-
 2087 cized as linguistically naïve (Kilgarriff, 1997). One way of testing this question is to ask
 2088 whether people tend to agree on the appropriate sense for example sentences: accord-
 2089 ing to Mihalcea et al. (2004), people agree on roughly 70% of examples using WordNet
 2090 senses; far better than chance, but less than agreement on other tasks, such as sentiment
 2091 annotation (Wilson et al., 2005).

2092 ***Other lexical semantic relations** Besides **synonymy**, WordNet also describes many
 2093 other lexical semantic relationships, including:

- 2094 • **antonymy**: *x* means the opposite of *y*, e.g. FRIEND-ENEMY;

³Navigli (2009) provides a survey of approaches for word-sense disambiguation.

⁴Several of the examples are adapted from WORDNET (Fellbaum, 2010).

- 2095 • **hyponymy:** x is a special case of y , e.g. RED-COLOR; the inverse relationship is
2096 **hyperonymy**;
- 2097 • **meronymy:** x is a part of y , e.g., WHEEL-BICYCLE; the inverse relationship is **holonymy**.

2098 Classification of these relations can be performed by searching for characteristic pat-
2099 terns between pairs of words, e.g., X , *such as* Y , which signals hyponymy (Hearst, 1992),
2100 or X *but* Y , which signals antonymy (Hatzivassiloglou and McKeown, 1997). Another ap-
2101 proach is to analyze each term's **distributional statistics** (the frequency of its neighboring
2102 words). Such approaches are described in detail in chapter 14.

2103 4.2.2 Word sense disambiguation as classification

2104 How can we tell living *plants* from manufacturing *plants*? The context is often critical:

- 2105 (4.4) a. Town officials are hoping to attract new manufacturing plants through weak-
2106 ened environmental regulations.
2107 b. The endangered plants play an important role in the local ecosystem.

It is possible to build a feature vector using the bag-of-words representation, by treat-
ing each context as a pseudo-document. The feature function is then,

$$f((plant, \text{The endangered plants play an ...}), y) = \\ \{(the, y) : 1, (\text{endangered}, y) : 1, (play, y) : 1, (an, y) : 1, \dots\}$$

2108 As in document classification, many of these features are irrelevant, but a few are very
2109 strong predictors. In this example, the context word *endangered* is a strong signal that
2110 the intended sense is biology rather than manufacturing. We would therefore expect a
2111 learning algorithm to assign high weight to (*endangered*, BIOLOGY), and low weight to
2112 (*endangered*, MANUFACTURING).⁵

It may also be helpful to go beyond the bag-of-words: for example, one might encode
the position of each context word with respect to the target, e.g.,

$$f((bank, I \text{ went to the bank to deposit my paycheck}), y) = \\ \{(i - 3, \text{went}, y) : 1, (i + 2, \text{deposit}, y) : 1, (i + 4, \text{paycheck}, y) : 1\}$$

2113 These are called **collocation features**, and they give more information about the specific
2114 role played by each context word. This idea can be taken further by incorporating addi-
2115 tional syntactic information about the grammatical role played by each context feature,
2116 such as the **dependency path** (see chapter 11).

⁵The context bag-of-words can be also used to perform word-sense disambiguation without machine learning: the Lesk (1986) algorithm selects the word sense whose dictionary definition best overlaps the local context.

Using such features, a classifier can be trained from labeled data. A **semantic concordance** is a corpus in which each open-class word (nouns, verbs, adjectives, and adverbs) is tagged with its word sense from the target dictionary or thesaurus. SemCor is a semantic concordance built from 234K tokens of the Brown corpus (Francis and Kucera, 1982), annotated as part of the WORDNET project (Fellbaum, 2010). SemCor annotations look like this:

(4.5) As of Sunday¹_N night¹_N there was⁴_V no word²_N ...,

with the superscripts indicating the annotated sense of each polysemous word, and the subscripts indicating the part-of-speech.

As always, supervised classification is only possible if enough labeled examples can be accumulated. This is difficult in word sense disambiguation, because each polysemous lemma requires its own training set: having a good classifier for the senses of *serve* is no help towards disambiguating *plant*. For this reason, unsupervised and **semi-supervised** methods are particularly important for word sense disambiguation (e.g., Yarowsky, 1995). These methods will be discussed in chapter 5. Unsupervised methods typically lean on the heuristic of “one sense per discourse”, which means that a lemma will usually have a single, consistent sense throughout any given document (Gale et al., 1992). Based on this heuristic, we can propagate information from high-confidence instances to lower-confidence instances in the same document (Yarowsky, 1995). Semi-supervised methods combine labeled and unlabeled data, and are discussed in more detail in chapter 5.

4.3 Design decisions for text classification

Text classification involves a number of design decisions. In some cases, the design decision is clear from the mathematics: if you are using regularization, then a regularization weight λ must be chosen. Other decisions are more subtle, arising only in the low level “plumbing” code that ingests and processes the raw data. Such decision can be surprisingly consequential for classification accuracy.

4.3.1 What is a word?

The bag-of-words representation presupposes that extracting a vector of word counts from text is unambiguous. But text documents are generally represented as a sequences of characters (in an encoding such as ascii or unicode), and the conversion to bag-of-words presupposes a definition of the “words” that are to be counted.

Whitespace	Isn't Ahab, Ahab? ;)
Treebank	Is n't Ahab , Ahab ? ;)
Tweet	Isn't Ahab , Ahab ? ;)
TokTok (Dehdari, 2014)	Isn ' t Ahab , Ahab ? ;)

Figure 4.1: The output of four NLTK tokenizers, applied to the string *Isn't Ahab, Ahab? ;)*

2148 **Tokenization**

2149 The first subtask for constructing a bag-of-words vector is **tokenization**: converting the
 2150 text from a sequence of characters to a sequence of **word!tokens**. A simple approach is
 2151 to define a subset of characters as whitespace, and then split the text on these tokens.
 2152 However, whitespace-based tokenization is not ideal: we may want to split conjunctions
 2153 like *isn't* and hyphenated phrases like *prize-winning* and *half-asleep*, and we likely want
 2154 to separate words from commas and periods that immediately follow them. At the same
 2155 time, it would be better not to split abbreviations like *U.S.* and *Ph.D.* In languages with
 2156 Roman scripts, tokenization is typically performed using regular expressions, with mod-
 2157 ules designed to handle each of these cases. For example, the NLTK package includes a
 2158 number of tokenizers (Loper and Bird, 2002); the outputs of four of the better-known tok-
 2159 enizers are shown in Figure 4.1. Social media researchers have found that emoticons and
 2160 other forms of orthographic variation pose new challenges for tokenization, leading to the
 2161 development of special purpose tokenizers to handle these phenomena (O'Connor et al.,
 2162 2010).

2163 Tokenization is a language-specific problem, and each language poses unique chal-
 2164 lenges. For example, Chinese does not include spaces between words, nor any other
 2165 consistent orthographic markers of word boundaries. A “greedy” approach is to scan the
 2166 input for character substrings that are in a predefined lexicon. However, Xue et al. (2003)
 2167 notes that this can be ambiguous, since many character sequences could be segmented in
 2168 multiple ways. Instead, he trains a classifier to determine whether each Chinese character,
 2169 or **hanzi**, is a word boundary. More advanced sequence labeling methods for word seg-
 2170 mentation are discussed in § 8.4. Similar problems can occur in languages with alphabetic
 2171 scripts, such as German, which does not include whitespace in compound nouns, yield-
 2172 ing examples such as *Freundschaftsbezeugungen* (demonstration of friendship) and *Dilett-*
 2173 *tantenaufdringlichkeiten* (the importunities of dilettantes). As Twain (1997) argues, “*These*
 2174 *things are not words, they are alphabetic processions.*” Social media raises similar problems
 2175 for English and other languages, with hashtags such as *#TrueLoveInFourWords* requiring
 2176 decomposition for analysis (Brun and Roux, 2014).

Original	The	Williams	sisters	are	leaving	this	tennis	centre
Porter stemmer	the	william	sister	are	leav	thi	tenni	centr
Lancaster stemmer	the	william	sist	ar	leav	thi	ten	cent
WordNet lemmatizer	The	Williams	sister	are	leaving	this	tennis	centre

Figure 4.2: Sample outputs of the Porter (1980) and Lancaster (Paice, 1990) stemmers, and the WORDNET lemmatizer

2177 Text normalization

2178 After splitting the text into tokens, the next question is which tokens are really distinct.
 2179 Is it necessary to distinguish *great*, *Great*, and *GREAT*? Sentence-initial capitalization may
 2180 be irrelevant to the classification task. Going further, the complete elimination of case
 2181 distinctions will result in a smaller vocabulary, and thus smaller feature vectors. However,
 2182 case distinctions might be relevant in some situations: for example, *apple* is a delicious
 2183 pie filling, while *Apple* is a company that specializes in proprietary dongles and power
 2184 adapters.

2185 For Roman script, case conversion can be performed using unicode string libraries.
 2186 Many scripts do not have case distinctions (e.g., the Devanagari script used for South
 2187 Asian languages, the Thai alphabet, and Japanese kana), and case conversion for all scripts
 2188 may not be available in every programming environment. (Unicode support is an im-
 2189 portant distinction between Python’s versions 2 and 3, and is a good reason for mi-
 2190 grating to Python 3 if you have not already done so. Compare the output of the code
 2191 "`\à l\’hôtel`".upper() in the two language versions.)

2192 Case conversion is a type of **text normalization**, which refers to string transfor-
 2193 mations that remove distinctions that are irrelevant to downstream applications (Sproat et al.,
 2194 2001). Other forms of normalization include the standardization of numbers (e.g., 1,000 to
 2195 1000) and dates (e.g., *August 11, 2015* to *2015/11/08*). Depending on the application, it may
 2196 even be worthwhile to convert all numbers and dates to special tokens, !NUM and !DATE.
 2197 In social media, there are additional orthographic phenomena that may be normalized,
 2198 such as expressive lengthening, e.g., *coooooool* (Aw et al., 2006; Yang and Eisenstein, 2013).
 2199 Similarly, historical texts feature spelling variations that may need to be normalized to a
 2200 contemporary standard form (Baron and Rayson, 2008).

2201 A more extreme form of normalization is to eliminate **inflectional affixes**, such as the
 2202 *-ed* and *-s* suffixes in English. On this view, *bike*, *bikes*, *biking*, and *biked* all refer to the
 2203 same underlying concept, so they should be grouped into a single feature. A **stemmer** is
 2204 a program for eliminating affixes, usually by applying a series of regular expression sub-
 2205 stitutions. Character-based stemming algorithms are necessarily approximate, as shown
 2206 in Figure 4.2: the Lancaster stemmer incorrectly identifies *-ers* as an inflectional suffix of

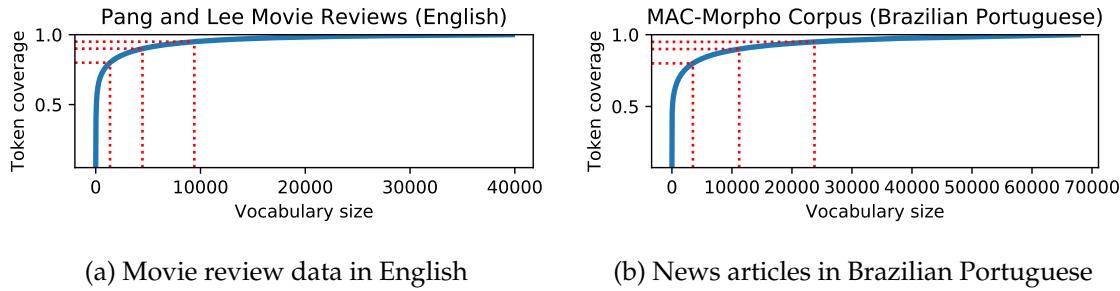


Figure 4.3: Tradeoff between token coverage (y-axis) and vocabulary size, on the NLTK movie review dataset, after sorting the vocabulary by decreasing frequency. The red dashed lines indicate 80%, 90%, and 95% coverage.

2207 *sisters* (by analogy to *fix/fixers*), and both stemmers incorrectly identify *-s* as a suffix of *this*
 2208 and *Williams*. Fortunately, even inaccurate stemming can improve bag-of-words classifi-
 2209 cation models, by merging related strings and thereby reducing the vocabulary size.

2210 Accurately handling irregular orthography requires word-specific rules. **Lemmatizers**
 2211 are systems that identify the underlying lemma of a given wordform. They must avoid the
 2212 over-generalization errors of the stemmers in Figure 4.2, and also handle more complex
 2213 transformations, such as *geese*→*goose*. The output of the WordNet lemmatizer is shown in
 2214 the final line of Figure 4.2. Both stemming and lemmatization are language-specific: an
 2215 English stemmer or lemmatizer is of little use on a text written in another language. The
 2216 discipline of **morphology** relates to the study of word-internal structure, and is described
 2217 in more detail in § 9.1.2.

2218 The value of normalization depends on the data and the task. Normalization re-
 2219 duces the size of the feature space, which can help in generalization. However, there
 2220 is always the risk of merging away linguistically meaningful distinctions. In supervised
 2221 machine learning, regularization and smoothing can play a similar role to normalization
 2222 — preventing the learner from overfitting to rare features — while avoiding the language-
 2223 specific engineering required for accurate normalization. In unsupervised scenarios, such
 2224 as content-based information retrieval (Manning et al., 2008) and topic modeling (Blei
 2225 et al., 2003), normalization is more critical.

2226 4.3.2 How many words?

2227 Limiting the size of the feature vector reduces the memory footprint of the resulting mod-
 2228 els, and increases the speed of prediction. Normalization can help to play this role, but
 2229 a more direct approach is simply to limit the vocabulary to the N most frequent words
 2230 in the dataset. For example, in the MOVIE-REVIEWS dataset provided with NLTK (origi-
 2231 nally from Pang et al., 2002), there are 39,768 word types, and 1.58M tokens. As shown

in Figure 4.3a, the most frequent 4000 word types cover 90% of all tokens, offering an order-of-magnitude reduction in the model size. Such ratios are language-specific: in for example, in the Brazilian Portuguese Mac-Morpho corpus (Aluísio et al., 2003), attaining 90% coverage requires more than 10000 word types (Figure 4.3b). This reflects the morphological complexity of Portuguese, which includes many more inflectional suffixes than English.

Eliminating rare words is not always advantageous for classification performance: for example, names, which are typically rare, play a large role in distinguishing topics of news articles. Another way to reduce the size of the feature space is to eliminate **stopwords** such as *the*, *to*, and *and*, which may seem to play little role in expressing the topic, sentiment, or stance. This is typically done by creating a **stoplist** (e.g., NLTK.CORPUS.STOPWORDS), and then ignoring all terms that match the list. However, corpus linguists and social psychologists have shown that seemingly inconsequential words can offer surprising insights about the author or nature of the text (Biber, 1991; Chung and Pennebaker, 2007). Furthermore, high-frequency words are unlikely to cause overfitting in discriminative classifiers. As with normalization, stopword filtering is more important for unsupervised problems, such as term-based document retrieval.

Another alternative for controlling model size is **feature hashing** (Weinberger et al., 2009). Each feature is assigned an index using a hash function. If a hash function that permits collisions is chosen (typically by taking the hash output modulo some integer), then the model can be made arbitrarily small, as multiple features share a single weight. Because most features are rare, accuracy is surprisingly robust to such collisions (Ganchev and Dredze, 2008).

4.3.3 Count or binary?

Finally, we may consider whether we want our feature vector to include the *count* of each word, or its *presence*. This gets at a subtle limitation of linear classification: it's worse to have two *failures* than one, but is it really twice as bad? Motivated by this intuition, Pang et al. (2002) use binary indicators of presence or absence in the feature vector: $f_j(x, y) \in \{0, 1\}$. They find that classifiers trained on these binary vectors tend to outperform feature vectors based on word counts. One explanation is that words tend to appear in clumps: if a word has appeared once in a document, it is likely to appear again (Church, 2000). These subsequent appearances can be attributed to this tendency towards repetition, and thus provide little additional information about the class label of the document.

4.4 Evaluating classifiers

In any supervised machine learning application, it is critical to reserve a held-out test set. This data should be used for only one purpose: to evaluate the overall accuracy of a single

2268 classifier. Using this data more than once would cause the estimated accuracy to be overly
 2269 optimistic, because the classifier would be customized to this data, and would not perform
 2270 as well as on unseen data in the future. It is usually necessary to set hyperparameters or
 2271 perform feature selection, so you may need to construct a **tuning** or **development set** for
 2272 this purpose, as discussed in § 2.2.5.

2273 There are a number of ways to evaluate classifier performance. The simplest is **accu-**
 2274 **racy**: the number of correct predictions, divided by the total number of instances,

$$\text{acc}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{N} \sum_i \delta(y^{(i)} = \hat{y}). \quad [4.4]$$

2275 Exams are usually graded by accuracy. Why are other metrics necessary? The main
 2276 reason is **class imbalance**. Suppose you are building a classifier to detect whether an
 2277 electronic health record (EHR) describes symptoms of a rare disease, which appears in
 2278 only 1% of all documents in the dataset. A classifier that reports $\hat{y} = \text{NEGATIVE}$ for
 2279 all documents would achieve 99% accuracy, but would be practically useless. We need
 2280 metrics that are capable of detecting the classifier's ability to discriminate between classes,
 2281 even when the distribution is skewed.

2282 One solution is to build a **balanced test set**, in which each possible label is equally rep-
 2283 resented. But in the EHR example, this would mean throwing away 98% of the original
 2284 dataset! Furthermore, the detection threshold itself might be a design consideration: in
 2285 health-related applications, we might prefer a very sensitive classifier, which returned a
 2286 positive prediction if there is even a small chance that $y^{(i)} = \text{POSITIVE}$. In other applica-
 2287 tions, a positive result might trigger a costly action, so we would prefer a classifier that
 2288 only makes positive predictions when absolutely certain. We need additional metrics to
 2289 capture these characteristics.

2290 4.4.1 Precision, recall, and *F*-MEASURE

2291 For any label (e.g., positive for presence of symptoms of a disease), there are two possible
 2292 errors:

- 2293 • **False positive**: the system incorrectly predicts the label.
- 2294 • **False negative**: the system incorrectly fails to predict the label.

2295 Similarly, for any label, there are two ways to be correct:

- 2296 • **True positive**: the system correctly predicts the label.
- 2297 • **True negative**: the system correctly predicts that the label does not apply to this
 2298 instance.

Classifiers that make a lot of false positives have low **precision**: they predict the label even when it isn't there. Classifiers that make a lot of false negatives have low **recall**: they fail to predict the label, even when it is there. These metrics distinguish these two sources of error, and are defined formally as:

$$\text{RECALL}(\mathbf{y}, \hat{\mathbf{y}}, k) = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad [4.5]$$

$$\text{PRECISION}(\mathbf{y}, \hat{\mathbf{y}}, k) = \frac{\text{TP}}{\text{TP} + \text{FP}}. \quad [4.6]$$

2299 Recall and precision are both conditional likelihoods of a correct prediction, which is why
 2300 their numerators are the same. Recall is conditioned on k being the correct label, $y^{(i)} = k$,
 2301 so the denominator sums over true positive and false negatives. Precision is conditioned
 2302 on k being the prediction, so the denominator sums over true positives and false positives.
 2303 Note that true negatives are not considered in either statistic. The classifier that labels
 2304 every document as "negative" would achieve zero recall; precision would be $\frac{0}{0}$.

2305 Recall and precision are complementary. A high-recall classifier is preferred when
 2306 false positives are cheaper than false negatives: for example, in a preliminary screening
 2307 for symptoms of a disease, the cost of a false positive might be an additional test, while a
 2308 false negative would result in the disease going untreated. Conversely, a high-precision
 2309 classifier is preferred when false positives are more expensive: for example, in spam de-
 2310 tection, a false negative is a relatively minor inconvenience, while a false positive might
 2311 mean that an important message goes unread.

The ***F*-MEASURE** combines recall and precision into a single metric, using the harmonic mean:

$$F\text{-MEASURE}(\mathbf{y}, \hat{\mathbf{y}}, k) = \frac{2rp}{r + p}, \quad [4.7]$$

2312 where r is recall and p is precision.⁶

Evaluating multi-class classification Recall, precision, and ***F*-MEASURE** are defined with respect to a specific label k . When there are multiple labels of interest (e.g., in word sense disambiguation or emotion classification), it is necessary to combine the ***F*-MEASURE** across each class. **Macro *F*-MEASURE** is the average ***F*-MEASURE** across several classes,

$$\text{Macro-}F(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} F\text{-MEASURE}(\mathbf{y}, \hat{\mathbf{y}}, k) \quad [4.8]$$

⁶ F -MEASURE is sometimes called F_1 , and generalizes to $F_\beta = \frac{(1+\beta^2)rp}{\beta^2p+r}$. The β parameter can be tuned to emphasize recall or precision.

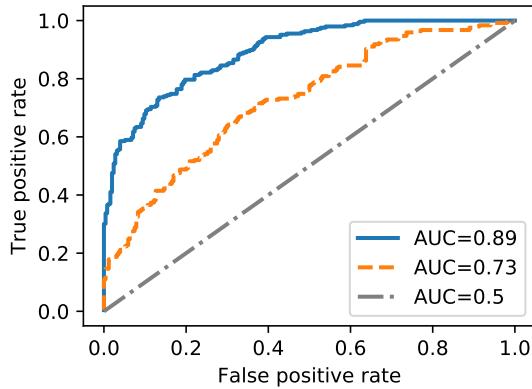


Figure 4.4: ROC curves for three classifiers of varying discriminative power, measured by AUC (area under the curve)

2313 In multi-class problems with unbalanced class distributions, the macro *F*-MEASURE is a
 2314 balanced measure of how well the classifier recognizes each class. In **micro F-MEASURE**,
 2315 we compute true positives, false positives, and false negatives for each class, and then add
 2316 them up to compute a single recall, precision, and *F*-MEASURE. This metric is balanced
 2317 across instances rather than classes, so it weights each class in proportion to its frequency
 2318 — unlike macro *F*-MEASURE, which weights each class equally.

2319 **4.4.2 Threshold-free metrics**

2320 In binary classification problems, it is possible to trade off between recall and precision by
 2321 adding a constant “threshold” to the output of the scoring function. This makes it possible
 2322 to trace out a curve, where each point indicates the performance at a single threshold. In
 2323 the **receiver operating characteristic (ROC)** curve,⁷ the *x*-axis indicates the **false positive**
 2324 **rate**, $\frac{FP}{FP+TN}$, and the *y*-axis indicates the recall, or **true positive rate**. A perfect classifier
 2325 attains perfect recall without any false positives, tracing a “curve” from the origin (0,0) to
 2326 the upper left corner (0,1), and then to (1,1). In expectation, a non-discriminative classifier
 2327 traces a diagonal line from the origin (0,0) to the upper right corner (1,1). Real classifiers
 2328 tend to fall between these two extremes. Examples are shown in Figure 4.4.

2329 The ROC curve can be summarized in a single number by taking its integral, the **area**
 2330 **under the curve (AUC)**. The AUC can be interpreted as the probability that a randomly-
 2331 selected positive example will be assigned a higher score by the classifier than a randomly-

⁷The name “receiver operator characteristic” comes from the metric’s origin in signal processing applications (Peterson et al., 1954). Other threshold-free metrics include **precision-recall curves**, **precision-at-*k***, and **balanced F-MEASURE**; see Manning et al. (2008) for more details.

selected negative example. A perfect classifier has $AUC = 1$ (all positive examples score higher than all negative examples); a non-discriminative classifier has $AUC = 0.5$ (given a randomly selected positive and negative example, either could score higher with equal probability); a perfectly wrong classifier would have $AUC = 0$ (all negative examples score higher than all positive examples). One advantage of AUC in comparison to F -MEASURE is that the baseline rate of 0.5 does not depend on the label distribution.

4.4.3 Classifier comparison and statistical significance

Natural language processing research and engineering often involves comparing different classification techniques. In some cases, the comparison is between algorithms, such as logistic regression versus averaged perceptron, or L_2 regularization versus L_1 . In other cases, the comparison is between feature sets, such as the bag-of-words versus positional bag-of-words (see § 4.2.2). **Ablation testing** involves systematically removing (ablating) various aspects of the classifier, such as feature groups, and testing the **null hypothesis** that the ablated classifier is as good as the full model.

A full treatment of hypothesis testing is beyond the scope of this text, but this section contains a brief summary of the techniques necessary to compare classifiers. The main aim of hypothesis testing is to determine whether the difference between two statistics — for example, the accuracies of two classifiers — is likely to arise by chance. We will be concerned with chance fluctuations that arise due to the finite size of the test set.⁸ An improvement of 10% on a test set with ten instances may reflect a random fluctuation that makes the test set more favorable to classifier c_1 than c_2 ; on another test set with a different ten instances, we might find that c_2 does better than c_1 . But if we observe the same 10% improvement on a test set with 1000 instances, this is highly unlikely to be explained by chance. Such a finding is said to be **statistically significant** at a level p , which is the probability of observing an effect of equal or greater magnitude when the null hypothesis is true. The notation $p < .05$ indicates that the likelihood of an equal or greater effect is less than 5%, assuming the null hypothesis is true.⁹

The binomial test

The statistical significance of a difference in accuracy can be evaluated using classical tests, such as the **binomial test**.¹⁰ Suppose that classifiers c_1 and c_2 disagree on N instances in a

⁸Other sources of variance include the initialization of non-convex classifiers such as neural networks, and the ordering of instances in online learning such as stochastic gradient descent and perceptron.

⁹Statistical hypothesis testing is useful only to the extent that the existing test set is representative of the instances that will be encountered in the future. If, for example, the test set is constructed from news documents, no hypothesis test can predict which classifier will perform best on documents from another domain, such as electronic health records.

¹⁰A well-known alternative to the binomial test is **McNemar's test**, which computes a **test statistic** based on the number of examples that are correctly classified by one system and incorrectly classified by the other.

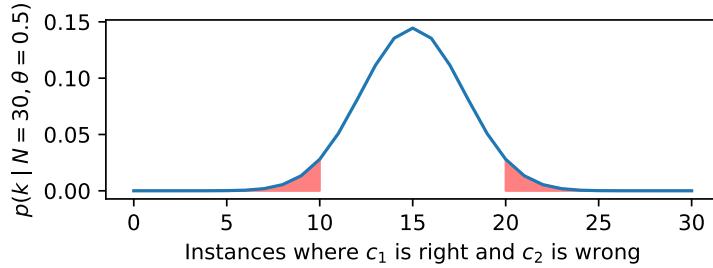


Figure 4.5: Probability mass function for the binomial distribution. The pink highlighted areas represent the cumulative probability for a significance test on an observation of $k = 10$ and $N = 30$.

2362 test set with binary labels, and that c_1 is correct on k of those instances. Under the null hy-
 2363 pothesis that the classifiers are equally accurate, we would expect k/N to be roughly equal
 2364 to $1/2$, and as N increases, k/N should be increasingly close to this expected value. These
 2365 properties are captured by the **binomial distribution**, which is a probability over counts
 2366 of binary random variables. We write $k \sim \text{Binom}(\theta, N)$ to indicate that k is drawn from
 2367 a binomial distribution, with parameter N indicating the number of random “draws”,
 2368 and θ indicating the probability of “success” on each draw. Each draw is an example on
 2369 which the two classifiers disagree, and a “success” is a case in which c_1 is right and c_2 is
 2370 wrong. (The label space is assumed to be binary, so if the classifiers disagree, exactly one
 2371 of them is correct. The test can be generalized to multi-class classification by focusing on
 2372 the examples in which exactly one classifier is correct.)

2373 The **probability mass function** (PMF) of the binomial distribution is,

$$p_{\text{Binom}}(k; N, \theta) = \binom{N}{k} \theta^k (1 - \theta)^{N-k}, \quad [4.9]$$

2374 with θ^k representing the probability of the k successes, $(1 - \theta)^{N-k}$ representing the prob-
 2375 ability of the $N - k$ unsuccessful draws. The expression $\binom{N}{k} = \frac{N!}{k!(N-k)!}$ is a binomial
 2376 coefficient, representing the number of possible orderings of events; this ensures that the
 2377 distribution sums to one over all $k \in \{0, 1, 2, \dots, N\}$.

Under the null hypothesis, when the classifiers disagree, each classifier is equally likely to be right, so $\theta = \frac{1}{2}$. Now suppose that among N disagreements, c_1 is correct $k < \frac{N}{2}$ times. The probability of c_1 being correct k or fewer times is the **one-tailed p -value**,

The null hypothesis distribution for this test statistic is known to be drawn from a chi-squared distribution with a single degree of freedom, so a p -value can be computed from the cumulative density function of this distribution (Dietterich, 1998). Both tests give similar results in most circumstances, but the binomial test is easier to understand from first principles.

because it is computed from the area under the binomial probability mass function from 0 to k , as shown in the left tail of Figure 4.5. This **cumulative probability** is computed as a sum over all values $i \leq k$,

$$\Pr_{\text{Binom}} \left(\text{count}(\hat{y}_2^{(i)}) = y^{(i)} \neq \hat{y}_1^{(i)} \right) \leq k; N, \theta = \frac{1}{2} = \sum_{i=0}^k p_{\text{Binom}} \left(i; N, \theta = \frac{1}{2} \right). \quad [4.10]$$

2378 The one-tailed p-value applies only to the asymmetric null hypothesis that c_1 is at least
 2379 as accurate as c_2 . To test the **two-tailed** null hypothesis that c_1 and c_2 are equally accu-
 2380 rate, we would take the sum of one-tailed p-values, where the second term is computed
 2381 from the right tail of Figure 4.5. The binomial distribution is symmetric, so this can be
 2382 computed by simply doubling the one-tailed p-value.

2383 Two-tailed tests are more stringent, but they are necessary in cases in which there is
 2384 no prior intuition about whether c_1 or c_2 is better. For example, in comparing logistic
 2385 regression versus averaged perceptron, a two-tailed test is appropriate. In an ablation
 2386 test, c_2 may contain a superset of the features available to c_1 . If the additional features are
 2387 thought to be likely to improve performance, then a one-tailed test would be appropriate,
 2388 if chosen in advance. However, such a test can only prove that c_2 is more accurate than
 2389 c_1 , and not the reverse.

2390 *Randomized testing

2391 The binomial test is appropriate for accuracy, but not for more complex metrics such as
 2392 *F*-MEASURE. To compute statistical significance for arbitrary metrics, we can apply ran-
 2393 domization. Specifically, draw a set of M **bootstrap samples** (Efron and Tibshirani, 1993),
 2394 by resampling instances from the original test set with replacement. Each bootstrap sam-
 2395 ple is itself a test set of size N . Some instances from the original test set will not appear
 2396 in any given bootstrap sample, while others will appear multiple times; but overall, the
 2397 sample will be drawn from the same distribution as the original test set. We can then com-
 2398 pute any desired evaluation on each bootstrap sample, which gives a distribution over the
 2399 value of the metric. Algorithm 7 shows how to perform this computation.

2400 To compare the *F*-MEASURE of two classifiers c_1 and c_2 , we set the function $\delta(\cdot)$ to
 2401 compute the difference in *F*-MEASURE on the bootstrap sample. If the difference is less
 2402 than or equal to zero in at least 5% of the samples, then we cannot reject the one-tailed
 2403 null hypothesis that c_2 is at least as good as c_1 (Berg-Kirkpatrick et al., 2012). We may
 2404 also be interested in the 95% **confidence interval** around a metric of interest, such as
 2405 the *F*-MEASURE of a single classifier. This can be computed by sorting the output of
 2406 Algorithm 7, and then setting the top and bottom of the 95% confidence interval to the
 2407 values at the 2.5% and 97.5% percentiles of the sorted outputs. Alternatively, you can fit
 2408 a normal distribution to the set of differences across bootstrap samples, and compute a
 2409 Gaussian confidence interval from the mean and variance.

Algorithm 7 Bootstrap sampling for classifier evaluation. The original test set is $\{\mathbf{x}^{(1:N)}, \mathbf{y}^{(1:N)}\}$, the metric is $\delta(\cdot)$, and the number of samples is M .

```

procedure BOOTSTRAP-SAMPLE( $\mathbf{x}^{(1:N)}, \mathbf{y}^{(1:N)}, \delta(\cdot), M$ )
  for  $t \in \{1, 2, \dots, M\}$  do
    for  $i \in \{1, 2, \dots, N\}$  do
       $j \sim \text{UniformInteger}(1, N)$ 
       $\tilde{\mathbf{x}}^{(i)} \leftarrow \mathbf{x}^{(j)}$ 
       $\tilde{\mathbf{y}}^{(i)} \leftarrow \mathbf{y}^{(j)}$ 
       $d^{(t)} \leftarrow \delta(\tilde{\mathbf{x}}^{(1:N)}, \tilde{\mathbf{y}}^{(1:N)})$ 
  return  $\{d^{(t)}\}_{t=1}^M$ 
```

2410 As the number of bootstrap samples goes to infinity, $M \rightarrow \infty$, the bootstrap estimate
 2411 is increasingly accurate. A typical choice for M is 10^4 or 10^5 ; larger numbers of samples
 2412 are necessary for smaller p -values. One way to validate your choice of M is to run the test
 2413 multiple times, and ensure that the p -values are similar; if not, increase M by an order of
 2414 magnitude. This is a heuristic measure of the **variance** of the test, which can decrease
 2415 with the square root \sqrt{M} (Robert and Casella, 2013).

2416 **4.4.4 *Multiple comparisons**

2417 Sometimes it is necessary to perform multiple hypothesis tests, such as when comparing
 2418 the performance of several classifiers on multiple datasets. Suppose you have five
 2419 datasets, and you compare four versions of your classifier against a baseline system, for a
 2420 total of 20 comparisons. Even if none of your classifiers is better than the baseline, there
 2421 will be some chance variation in the results, and in expectation you will get one statistically
 2422 significant improvement at $p = 0.05 = \frac{1}{20}$. It is therefore necessary to adjust the
 2423 p -values when reporting the results of multiple comparisons.

2424 One approach is to require a threshold of $\frac{\alpha}{m}$ to report a p value of $p < \alpha$ when performing m tests. This is known as the **Bonferroni correction**, and it limits the overall probability of incorrectly rejecting the null hypothesis at α . Another approach is to bound the **false discovery rate** (FDR), which is the fraction of null hypothesis rejections that are incorrect. Benjamini and Hochberg (1995) propose a p -value correction that bounds the fraction of false discoveries at α : sort the p -values of each individual test in ascending order, and set the significance threshold equal to largest k such that $p_k \leq \frac{k}{m}\alpha$. If $k > 1$, the FDR adjustment is more permissive than the Bonferroni correction.

2432 **4.5 Building datasets**

2433 Sometimes, if you want to build a classifier, you must first build a dataset of your own.
 2434 This includes selecting a set of documents or instances to annotate, and then performing
 2435 the annotations. The scope of the dataset may be determined by the application: if you
 2436 want to build a system to classify electronic health records, then you must work with a
 2437 corpus of records of the type that your classifier will encounter when deployed. In other
 2438 cases, the goal is to build a system that will work across a broad range of documents. In
 2439 this case, it is best to have a *balanced* corpus, with contributions from many styles and
 2440 genres. For example, the Brown corpus draws from texts ranging from government doc-
 2441 uments to romance novels (Francis, 1964), and the Google Web Treebank includes an-
 2442 notations for five “domains” of web documents: question answers, emails, newsgroups,
 2443 reviews, and blogs (Petrov and McDonald, 2012).

2444 **4.5.1 Metadata as labels**

2445 Annotation is difficult and time-consuming, and most people would rather avoid it. It
 2446 is sometimes possible to exploit existing metadata to obtain labels for training a classi-
 2447 fier. For example, reviews are often accompanied by a numerical rating, which can be
 2448 converted into a classification label (see § 4.1). Similarly, the nationalities of social media
 2449 users can be estimated from their profiles (Dredze et al., 2013) or even the time zones of
 2450 their posts (Gouws et al., 2011). More ambitiously, we may try to classify the political af-
 2451 filiations of social media profiles based on their social network connections to politicians
 2452 and major political parties (Rao et al., 2010).

2453 The convenience of quickly constructing large labeled datasets without manual an-
 2454 notation is appealing. However this approach relies on the assumption that unlabeled
 2455 instances — for which metadata is unavailable — will be similar to labeled instances.
 2456 Consider the example of labeling the political affiliation of social media users based on
 2457 their network ties to politicians. If a classifier attains high accuracy on such a test set,
 2458 is it safe to assume that it accurately predicts the political affiliation of all social media
 2459 users? Probably not. Social media users who establish social network ties to politicians
 2460 may be more likely to mention politics in the text of their messages, as compared to the
 2461 average user, for whom no political metadata is available. If so, the accuracy on a test set
 2462 constructed from social network metadata would give an overly optimistic picture of the
 2463 method’s true performance on unlabeled data.

2464 **4.5.2 Labeling data**

2465 In many cases, there is no way to get ground truth labels other than manual annotation.
 2466 An annotation protocol should satisfy several criteria: the annotations should be *expressive*
 2467 enough to capture the phenomenon of interest; they should be *replicable*, meaning that

2468 another annotator or team of annotators would produce very similar annotations if given
2469 the same data; and they should be *scalable*, so that they can be produced relatively quickly.
2470 Hovy and Lavid (2010) propose a structured procedure for obtaining annotations that
2471 meet these criteria, which is summarized below.

- 2472 1. **Determine what the annotations are to include.** This is usually based on some
2473 theory of the underlying phenomenon: for example, if the goal is to produce an-
2474 notations about the emotional state of a document’s author, one should start with a
2475 theoretical account of the types or dimensions of emotion (e.g., Mohammad and Tur-
2476 ney, 2013). At this stage, the tradeoff between expressiveness and scalability should
2477 be considered: a full instantiation of the underlying theory might be too costly to
2478 annotate at scale, so reasonable approximations should be considered.
- 2479 2. Optionally, one may **design or select a software tool to support the annotation**
2480 **effort.** Existing general-purpose annotation tools include BRAT (Stenetorp et al.,
2481 2012) and MMAX2 (Müller and Strube, 2006).
- 2482 3. **Formalize the instructions for the annotation task.** To the extent that the instruc-
2483 tions are not explicit, the resulting annotations will depend on the intuitions of the
2484 annotators. These intuitions may not be shared by other annotators, or by the users
2485 of the annotated data. Therefore explicit instructions are critical to ensuring the an-
2486 notations are replicable and usable by other researchers.
- 2487 4. **Perform a pilot annotation** of a small subset of data, with multiple annotators for
2488 each instance. This will give a preliminary assessment of both the replicability and
2489 scalability of the current annotation instructions. Metrics for computing the rate of
2490 agreement are described below. Manual analysis of specific disagreements should
2491 help to clarify the instructions, and may lead to modifications of the annotation task
2492 itself. For example, if two labels are commonly conflated by annotators, it may be
2493 best to merge them.
- 2494 5. **Annotate the data.** After finalizing the annotation protocol and instructions, the
2495 main annotation effort can begin. Some, if not all, of the instances should receive
2496 multiple annotations, so that inter-annotator agreement can be computed. In some
2497 annotation projects, instances receive many annotations, which are then aggregated
2498 into a “consensus” label (e.g., Danescu-Niculescu-Mizil et al., 2013). However, if the
2499 annotations are time-consuming or require significant expertise, it may be preferable
2500 to maximize scalability by obtaining multiple annotations for only a small subset of
2501 examples.
- 2502 6. **Compute and report inter-annotator agreement, and release the data.** In some
2503 cases, the raw text data cannot be released, due to concerns related to copyright or

privacy. In these cases, one solution is to publicly release **stand-off annotations**, which contain links to document identifiers. The documents themselves can be released under the terms of a licensing agreement, which can impose conditions on how the data is used. It is important to think through the potential consequences of releasing data: people may make personal data publicly available without realizing that it could be redistributed in a dataset and publicized far beyond their expectations (boyd and Crawford, 2012).

2511 Measuring inter-annotator agreement

2512 To measure the replicability of annotations, a standard practice is to compute the extent to
 2513 which annotators agree with each other. If the annotators frequently disagree, this casts
 2514 doubt on either their reliability or on the annotation system itself. For classification, one
 2515 can compute the frequency with which the annotators agree; for rating scales, one can
 2516 compute the average distance between ratings. These raw agreement statistics must then
 2517 be compared with the rate of agreement by chance — the expected level of agreement that
 2518 would be obtained between two annotators who ignored the data.

2519 **Cohen’s Kappa** is widely used for quantifying the agreement on discrete labeling
 2520 tasks (Cohen, 1960; Carletta, 1996),¹¹

$$\kappa = \frac{\text{agreement} - E[\text{agreement}]}{1 - E[\text{agreement}]}. \quad [4.11]$$

2521 The numerator is the difference between the observed agreement and the chance agree-
 2522 ment, and the denominator is the difference between perfect agreement and chance agree-
 2523 ment. Thus, $\kappa = 1$ when the annotators agree in every case, and $\kappa = 0$ when the annota-
 2524 tors agree only as often as would happen by chance. Various heuristic scales have been
 2525 proposed for determining when κ indicates “moderate”, “good”, or “substantial” agree-
 2526 ment; for reference, Lee and Narayanan (2005) report $\kappa \approx 0.45 - 0.47$ for annotations
 2527 of emotions in spoken dialogues, which they describe as “moderate agreement”; Stolcke
 2528 et al. (2000) report $\kappa = 0.8$ for annotations of **dialogue acts**, which are labels for the pur-
 2529 pose of each turn in a conversation.

2530 When there are two annotators, the expected chance agreement is computed as,

$$E[\text{agreement}] = \sum_k \hat{\Pr}(Y = k)^2, \quad [4.12]$$

2531 where k is a sum over labels, and $\hat{\Pr}(Y = k)$ is the empirical probability of label k across
 2532 all annotations. The formula is derived from the expected number of agreements if the
 2533 annotations were randomly shuffled. Thus, in a binary labeling task, if one label is applied
 2534 to 90% of instances, chance agreement is $.9^2 + .1^2 = .82$.

¹¹For other types of annotations, Krippendorff’s alpha is a popular choice (Hayes and Krippendorff, 2007; Artstein and Poesio, 2008).

2535 **Crowdsourcing**

2536 Crowdsourcing is often used to rapidly obtain annotations for classification problems.
 2537 For example, **Amazon Mechanical Turk** makes it possible to define “human intelligence
 2538 tasks (hits)”, such as labeling data. The researcher sets a price for each set of annotations
 2539 and a list of minimal qualifications for annotators, such as their native language and their
 2540 satisfaction rate on previous tasks. The use of relatively untrained “crowdworkers” con-
 2541 trasts with earlier annotation efforts, which relied on professional linguists (Marcus et al.,
 2542 1993). However, crowdsourcing has been found to produce reliable annotations for many
 2543 language-related tasks (Snow et al., 2008). Crowdsourcing is part of the broader field
 2544 of **human computation** (Law and Ahn, 2011). For a critical examination of ethical issues
 2545 related to crowdsourcing, see Fort et al. (2011).

2546 **Additional resources**

2547 Many of the preprocessing issues discussed in this chapter also arise in information re-
 2548 trieval. See Manning et al. (2008) for discussion of tokenization and related algorithms.

2549 **Exercises**

- 2550 1. As noted in § 4.3.3, words tend to appear in clumps, with subsequent occurrences
 2551 of a word being more probable. More concretely, if word j has probability $\phi_{y,j}$
 2552 of appearing in a document with label y , then the probability of two appearances
 2553 ($x_j^{(i)} = 2$) is greater than $\phi_{y,j}^2$.

2554 Suppose you are applying Naïve Bayes to a binary classification. Focus on a word j
 2555 which is more probable under label $y = 1$, so that,

$$\Pr(w = j \mid y = 1) > \Pr(w = j \mid y = 0). \quad [4.13]$$

2556 Now suppose that $x_j^{(i)} > 1$. All else equal, will the classifier overestimate or under-
 2557 estimate the posterior $\Pr(y = 1 \mid x)$?

- 2558 2. Prove that F-measure is never greater than the arithmetic mean of recall and pre-
 2559 cision, $\frac{r+p}{2}$. Your solution should also show that F-measure is equal to $\frac{r+p}{2}$ iff $r = p$.
- 2560 3. Given a binary classification problem in which the probability of the “positive” label
 2561 is equal to α , what is the expected F-MEASURE of a random classifier which ignores
 2562 the data, and selects $\hat{y} = +1$ with probability $\frac{1}{2}$? (Assume that $p(\hat{y}) \perp p(y)$.) What is
 2563 the expected F-MEASURE of a classifier that selects $\hat{y} = +1$ with probability α (also
 2564 independent of $y^{(i)}$)? Depending on α , which random classifier will score better?

- 2565 4. Suppose that binary classifiers c_1 and c_2 disagree on $N = 30$ cases, and that c_1 is
 2566 correct in $k = 10$ of those cases.
- 2567 • Write a program that uses primitive functions such as `exp` and `factorial` to com-
 2568 pute the **two-tailed** p -value — you may use an implementation of the “choose”
 2569 function if one is available. Verify your code against the output of a library for
 2570 computing the binomial test or the binomial CDF, such as `SCIPY.STATS.BINOM`
 2571 in Python.
- 2572 • Then use a randomized test to try to obtain the same p -value. In each sample,
 2573 draw from a binomial distribution with $N = 30$ and $\theta = \frac{1}{2}$. Count the fraction
 2574 of samples in which $k \leq 10$. This is the one-tailed p -value; double this to
 2575 compute the two-tailed p -value.
- 2576 • Try this with varying numbers of bootstrap samples: $M \in \{100, 1000, 5000, 10000\}$.
 2577 For $M = 100$ and $M = 1000$, run the test 10 times, and plot the resulting p -
 2578 values.
- 2579 • Finally, perform the same tests for $N = 70$ and $k = 25$.
- 2580 5. SemCor 3.0 is a labeled dataset for word sense disambiguation. You can download
 2581 it,¹² or access it in `NLTK.CORPORA.SEMCOR`.
- 2582 Choose a word that appears at least ten times in SemCor (*find*), and annotate its
 2583 WordNet senses across ten randomly-selected examples, without looking at the ground
 2584 truth. Use online WordNet to understand the definition of each of the senses.¹³ Have
 2585 a partner do the same annotations, and compute the raw rate of agreement, expected
 2586 chance rate of agreement, and Cohen’s kappa.
- 2587 6. Download the Pang and Lee movie review data, currently available from <http://www.cs.cornell.edu/people/pabo/movie-review-data/>. Hold out a
 2588 randomly-selected 400 reviews as a test set.
- 2589 Download a sentiment lexicon, such as the one currently available from Bing Liu,
 2590 <https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html>. Tokenize
 2591 the data, and classify each document as positive iff it has more positive sentiment
 2592 words than negative sentiment words. Compute the accuracy and *F*-MEASURE on
 2593 detecting positive reviews on the test set, using this lexicon-based classifier.
- 2594 Then train a discriminative classifier (averaged perceptron or logistic regression) on
 2595 the training set, and compute its accuracy and *F*-MEASURE on the test set.
- 2596 Determine whether the differences are statistically significant, using two-tailed hy-
 2597 pothesis tests: Binomial for the difference in accuracy, and bootstrap for the differ-
 2598 ence in macro-*F*-MEASURE.

¹²e.g., https://github.com/google-research-datasets/word_sense_disambiguation_corpora or <http://globalwordnet.org/wordnet-annotated-corpora/>

¹³<http://wordnetweb.princeton.edu/perl/webwn>

2600 The remaining problems will require you to build a classifier and test its properties. Pick
2601 a multi-class text classification dataset that is not already tokenized. One example is a
2602 dataset of New York Times headlines and topics (Boydston, 2013).¹⁴ Divide your data
2603 into training (60%), development (20%), and test sets (20%), if no such division already
2604 exists. If your dataset is very large, you may want to focus on a few thousand instances at
2605 first.

2606 7. Compare various vocabulary sizes of 10^2 , 10^3 , 10^4 , 10^5 , using the most frequent words
2607 in each case (you may use any reasonable tokenizer). Train logistic regression clas-
2608 sifiers for each vocabulary size, and apply them to the development set. Plot the
2609 accuracy and Macro-*F*-MEASURE with the increasing vocabulary size. For each vo-
2610 cabulary size, tune the regularizer to maximize accuracy on a subset of data that is
2611 held out from the training set.

2612 8. Compare the following tokenization algorithms:

- 2613 • Whitespace, using a regular expression;
2614 • The Penn Treebank tokenizer from NLTK;
2615 • Splitting the input into non-overlapping five-character units, regardless of whites-
2616 pace or punctuation.

2617 Compute the token/type ratio for each tokenizer on the training data, and explain
2618 what you find. Train your classifier on each tokenized dataset, tuning the regularizer
2619 on a subset of data that is held out from the training data. Tokenize the development
2620 set, and report accuracy and Macro-*F*-MEASURE.

2621 9. Apply the Porter and Lancaster stemmers to the training set, using any reasonable
2622 tokenizer, and compute the token/type ratios. Train your classifier on the stemmed
2623 data, and compute the accuracy and Macro-*F*-MEASURE on stemmed development
2624 data, again using a held-out portion of the training data to tune the regularizer.

2625 10. Identify the best combination of vocabulary filtering, tokenization, and stemming
2626 from the previous three problems. Apply this preprocessing to the test set, and
2627 compute the test set accuracy and Macro-*F*-MEASURE. Compare against a baseline
2628 system that applies no vocabulary filtering, whitespace tokenization, and no stem-
2629 ming.

2630 Use the binomial test to determine whether your best-performing system is signifi-
2631 cantly more accurate than the baseline.

¹⁴Available as a CSV file at <http://www.amber-boydston.com/supplementary-information-for-making-the-news.html>. Use the field TOPIC_2DIGIT for this problem.

2632 Use the bootstrap test with $M = 10^4$ to determine whether your best-performing
2633 system achieves significantly higher macro-*F*-MEASURE.

2634 Chapter 5

2635 Learning without supervision

2636 So far, we have assumed the following setup:

- 2637 a **training set** where you get observations x and labels y ;
- 2638 a **test set** where you only get observations x .

2639 Without labeled data, is it possible to learn anything? This scenario is known as **unsu-**
2640 **pervised learning**, and we will see that indeed it is possible to learn about the underlying
2641 structure of unlabeled observations. This chapter will also explore some related scenarios:
2642 **semi-supervised learning**, in which only some instances are labeled, and **domain adap-**
2643 **tation**, in which the training data differs from the data on which the trained system will
2644 be deployed.

2645 5.1 Unsupervised learning

2646 To motivate unsupervised learning, consider the problem of word sense disambiguation
2647 (§ 4.2). The goal is to classify each instance of a word, such as *bank* into a sense,

- 2648 bank#1: a financial institution
- 2649 bank#2: the land bordering a river

2650 It is difficult to obtain sufficient training data for word sense disambiguation, because
2651 even a large corpus will contain only a few instances of all but the most common words.
2652 Is it possible to learn anything about these different senses without labeled data?

2653 Word sense disambiguation is usually performed using feature vectors constructed
2654 from the local context of the word to be disambiguated. For example, for the word

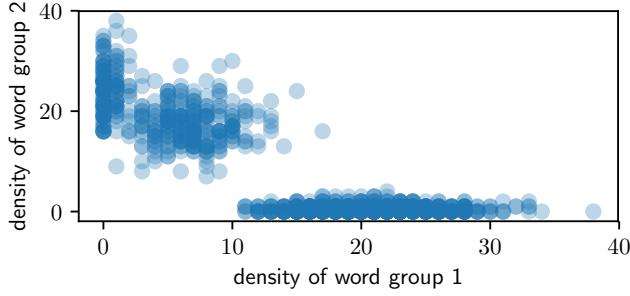


Figure 5.1: Counts of words from two different context groups

2655 *bank*, the immediate context might typically include words from one of the following two
 2656 groups:

- 2657 1. *financial, deposits, credit, lending, capital, markets, regulated, reserve, liquid, assets*
- 2658 2. *land, water, geography, stream, river, flow, deposits, discharge, channel, ecology*

2659 Now consider a scatterplot, in which each point is a document containing the word *bank*.
 2660 The location of the document on the x -axis is the count of words in group 1, and the
 2661 location on the y -axis is the count for group 2. In such a plot, shown in Figure 5.1, two
 2662 “blobs” might emerge, and these blobs correspond to the different senses of *bank*.

2663 Here’s a related scenario, from a different problem. Suppose you download thousands
 2664 of news articles, and make a scatterplot, where each point corresponds to a document:
 2665 the x -axis is the frequency of the group of words (*hurricane, winds, storm*); the y -axis is the
 2666 frequency of the group (*election, voters, vote*). This time, three blobs might emerge: one
 2667 for documents that are largely about a hurricane, another for documents largely about a
 2668 election, and a third for documents about neither topic.

2669 These clumps represent the underlying structure of the data. But the two-dimensional
 2670 scatter plots are based on groupings of context words, and in real scenarios these word
 2671 lists are unknown. Unsupervised learning applies the same basic idea, but in a high-
 2672 dimensional space with one dimension for every context word. This space can’t be di-
 2673 rectly visualized, but the goal is the same: try to identify the underlying structure of the
 2674 observed data, such that there are a few clusters of points, each of which is internally
 2675 coherent. **Clustering** algorithms are capable of finding such structure automatically.

2676 5.1.1 ***K*-means clustering**

2677 Clustering algorithms assign each data point to a discrete cluster, $z_i \in 1, 2, \dots, K$. One of
 2678 the best known clustering algorithms is ***K*-means**, an iterative algorithm that maintains

Algorithm 8 K -means clustering algorithm

```

1: procedure  $K$ -MEANS( $\mathbf{x}_{1:N}, K$ )
2:   for  $i \in 1 \dots N$  do                                 $\triangleright$  initialize cluster memberships
3:      $z^{(i)} \leftarrow \text{RANDOMINT}(1, K)$ 
4:   repeat
5:     for  $k \in 1 \dots K$  do                           $\triangleright$  recompute cluster centers
6:        $\boldsymbol{\nu}_k \leftarrow \frac{1}{\delta(z^{(i)}=k)} \sum_{i=1}^N \delta(z^{(i)} = k) \mathbf{x}^{(i)}$ 
7:     for  $i \in 1 \dots N$  do                       $\triangleright$  reassign instances to nearest clusters
8:        $z^{(i)} \leftarrow \text{argmin}_k \|\mathbf{x}^{(i)} - \boldsymbol{\nu}_k\|^2$ 
9:   until converged
10:  return  $\{z^{(i)}\}$                                  $\triangleright$  return cluster assignments

```

2679 a cluster assignment for each instance, and a central (“mean”) location for each cluster.
 2680 K -means iterates between updates to the assignments and the centers:

2681 1. each instance is placed in the cluster with the closest center;

2682 2. each center is recomputed as the average over points in the cluster.

2683 This procedure is formalized in Algorithm 8. The term $\|\mathbf{x}^{(i)} - \boldsymbol{\nu}\|^2$ refers to the squared
 2684 Euclidean norm, $\sum_{j=1}^V (x_j^{(i)} - \nu_j)^2$. An important property of K -means is that the con-
 2685 verged solution depends on the initialization, and a better clustering can sometimes be
 2686 found simply by re-running the algorithm from a different random starting point.

2687 **Soft K -means** is a particularly relevant variant. Instead of directly assigning each
 2688 point to a specific cluster, soft K -means assigns to each point a *distribution* over clusters
 2689 $\mathbf{q}^{(i)}$, so that $\sum_{k=1}^K q^{(i)}(k) = 1$, and $\forall_k, q^{(i)}(k) \geq 0$. The soft weight $q^{(i)}(k)$ is computed from
 2690 the distance of $\mathbf{x}^{(i)}$ to the cluster center $\boldsymbol{\nu}_k$. In turn, the center of each cluster is computed
 2691 from a weighted average of the points in the cluster,

$$\boldsymbol{\nu}_k = \frac{1}{\sum_{i=1}^N q^{(i)}(k)} \sum_{i=1}^N q^{(i)}(k) \mathbf{x}^{(i)}. \quad [5.1]$$

2692 We will now explore a probabilistic version of soft K -means clustering, based on **expectation-**
 2693 **maximization** (EM). Because EM clustering can be derived as an approximation to maximum-
 2694 likelihood estimation, it can be extended in a number of useful ways.

2695 5.1.2 Expectation-Maximization (EM)

Expectation-maximization combines the idea of soft K -means with Naïve Bayes classification. To review, Naïve Bayes defines a probability distribution over the data,

$$\log p(\mathbf{x}, \mathbf{y}; \boldsymbol{\phi}, \boldsymbol{\mu}) = \sum_{i=1}^N \log \left(p(\mathbf{x}^{(i)} | y^{(i)}; \boldsymbol{\phi}) \times p(y^{(i)}; \boldsymbol{\mu}) \right) \quad [5.2]$$

Now suppose that you never observe the labels. To indicate this, we'll refer to the label of each instance as $z^{(i)}$, rather than $y^{(i)}$, which is usually reserved for observed variables. By marginalizing over the **latent variables** z , we obtain the marginal probability of the observed instances \mathbf{x} :

$$\log p(\mathbf{x}; \boldsymbol{\phi}, \boldsymbol{\mu}) = \sum_{i=1}^N \log p(\mathbf{x}^{(i)}; \boldsymbol{\phi}, \boldsymbol{\mu}) \quad [5.3]$$

$$= \sum_{i=1}^N \log \sum_{z=1}^K p(\mathbf{x}^{(i)}, z; \boldsymbol{\phi}, \boldsymbol{\mu}) \quad [5.4]$$

$$= \sum_{i=1}^N \log \sum_{z=1}^K p(\mathbf{x}^{(i)} | z; \boldsymbol{\phi}) \times p(z; \boldsymbol{\mu}). \quad [5.5]$$

2696 The parameters $\boldsymbol{\phi}$ and $\boldsymbol{\mu}$ can be obtained by maximizing the marginal likelihood in
 2697 Equation 5.5. Why is this the right thing to maximize? Without labels, discriminative
 2698 learning is impossible — there's nothing to discriminate. So maximum likelihood is all
 2699 we have.

2700 When the labels are observed, we can estimate the parameters of the Naïve Bayes
 2701 probability model separately for each label. But marginalizing over the labels couples
 2702 these parameters, making direct optimization of $\log p(\mathbf{x})$ intractable. We will approxi-
 2703 mate the log-likelihood by introducing an auxiliary variable $\mathbf{q}^{(i)}$, which is a distribution
 2704 over the label set $\mathcal{Z} = \{1, 2, \dots, K\}$. The optimization procedure will alternate between
 2705 updates to \mathbf{q} and updates to the parameters $(\boldsymbol{\phi}, \boldsymbol{\mu})$. Thus, $\mathbf{q}^{(i)}$ plays here as in soft K -
 2706 means.

To derive the updates for this optimization, multiply the right side of Equation 5.5 by

the ratio $\frac{q^{(i)}(z)}{q^{(i)}(z)} = 1$,

$$\log p(\mathbf{x}; \phi, \mu) = \sum_{i=1}^N \log \sum_{z=1}^K p(\mathbf{x}^{(i)} | z; \phi) \times p(z; \mu) \times \frac{q^{(i)}(z)}{q^{(i)}(z)} \quad [5.6]$$

$$= \sum_{i=1}^N \log \sum_{z=1}^K q^{(i)}(z) \times p(\mathbf{x}^{(i)} | z; \phi) \times p(z; \mu) \times \frac{1}{q^{(i)}(z)} \quad [5.7]$$

$$= \sum_{i=1}^N \log E_{\mathbf{q}^{(i)}} \left[\frac{p(\mathbf{x}^{(i)} | z; \phi) p(z; \mu)}{q^{(i)}(z)} \right], \quad [5.8]$$

where $E_{\mathbf{q}^{(i)}} [f(z)] = \sum_{z=1}^K q^{(i)}(z) \times f(z)$ refers to the expectation of the function f under the distribution $z \sim \mathbf{q}^{(i)}$.

Jensen's inequality says that because \log is a concave function, we can push it inside the expectation, and obtain a lower bound.

$$\log p(\mathbf{x}; \phi, \mu) \geq \sum_{i=1}^N E_{\mathbf{q}^{(i)}} \left[\log \frac{p(\mathbf{x}^{(i)} | z; \phi) p(z; \mu)}{q^{(i)}(z)} \right] \quad [5.9]$$

$$J \triangleq \sum_{i=1}^N E_{\mathbf{q}^{(i)}} \left[\log p(\mathbf{x}^{(i)} | z; \phi) + \log p(z; \mu) - \log q^{(i)}(z) \right] \quad [5.10]$$

$$= \sum_{i=1}^N E_{\mathbf{q}^{(i)}} \left[\log p(\mathbf{x}^{(i)}, z; \phi, \mu) \right] + H(\mathbf{q}^{(i)}) \quad [5.11]$$

We will focus on Equation 5.10, which is the lower bound on the marginal log-likelihood of the observed data, $\log p(\mathbf{x})$. Equation 5.11 shows the connection to the information theoretic concept of **entropy**, $H(\mathbf{q}^{(i)}) = -\sum_{z=1}^K q^{(i)}(z) \log q^{(i)}(z)$, which measures the average amount of information produced by a draw from the distribution $q^{(i)}$. The lower bound J is a function of two groups of arguments:

- the distributions $\mathbf{q}^{(i)}$ for each instance;
- the parameters μ and ϕ .

The expectation-maximization (EM) algorithm maximizes the bound with respect to each of these arguments in turn, while holding the other fixed.

2718 The E-step

The step in which we update $\mathbf{q}^{(i)}$ is known as the **E-step**, because it updates the distribution under which the expectation is computed. To derive this update, first write out the

expectation in the lower bound as a sum,

$$J = \sum_{i=1}^N \sum_{z=1}^K q^{(i)}(z) \left[\log p(\mathbf{x}^{(i)} | z; \boldsymbol{\phi}) + \log p(z; \boldsymbol{\mu}) - \log q^{(i)}(z) \right]. \quad [5.12]$$

When optimizing this bound, we must also respect a set of “sum-to-one” constraints, $\sum_{z=1}^K q^{(i)}(z) = 1$ for all i . Just as in Naïve Bayes, this constraint can be incorporated into a Lagrangian:

$$J_q = \sum_{i=1}^N \sum_{z=1}^K q^{(i)}(z) \left(\log p(\mathbf{x}^{(i)} | z; \boldsymbol{\phi}) + \log p(z; \boldsymbol{\mu}) - \log q^{(i)}(z) \right) + \lambda^{(i)} \left(1 - \sum_{z=1}^K q^{(i)}(z) \right), \quad [5.13]$$

2719 where $\lambda^{(i)}$ is the Lagrange multiplier for instance i .

The Lagrangian is maximized by taking the derivative and solving for $q^{(i)}$:

$$\frac{\partial J_q}{\partial q^{(i)}(z)} = \log p(\mathbf{x}^{(i)} | z; \boldsymbol{\phi}) + \log p(z; \boldsymbol{\mu}) - \log q^{(i)}(z) - 1 - \lambda^{(i)} \quad [5.14]$$

$$\log q^{(i)}(z) = \log p(\mathbf{x}^{(i)} | z; \boldsymbol{\phi}) + \log p(z; \boldsymbol{\mu}) - 1 - \lambda^{(i)} \quad [5.15]$$

$$q^{(i)}(z) \propto p(\mathbf{x}^{(i)} | z; \boldsymbol{\phi}) \times p(z; \boldsymbol{\mu}). \quad [5.16]$$

Applying the sum-to-one constraint gives an exact solution,

$$q^{(i)}(z) = \frac{p(\mathbf{x}^{(i)} | z; \boldsymbol{\phi}) \times p(z; \boldsymbol{\mu})}{\sum_{z'=1}^K p(\mathbf{x}^{(i)} | z'; \boldsymbol{\phi}) \times p(z'; \boldsymbol{\mu})} \quad [5.17]$$

$$= p(z | \mathbf{x}^{(i)}; \boldsymbol{\phi}, \boldsymbol{\mu}). \quad [5.18]$$

2720 After normalizing, each $q^{(i)}$ — which is the soft distribution over clusters for data $\mathbf{x}^{(i)}$ —
2721 is set to the posterior probability $p(z | \mathbf{x}^{(i)}; \boldsymbol{\phi}, \boldsymbol{\mu})$ under the current parameters. Although
2722 the Lagrange multipliers $\lambda^{(i)}$ were introduced as additional parameters, they drop out
2723 during normalization.

2724 The M-step

Next, we hold fixed the soft assignments $q^{(i)}$, and maximize with respect to the parameters, $\boldsymbol{\phi}$ and $\boldsymbol{\mu}$. Let’s focus on the parameter $\boldsymbol{\phi}$, which parametrizes the likelihood $p(\mathbf{x} | z; \boldsymbol{\phi})$, and leave $\boldsymbol{\mu}$ for an exercise. The parameter $\boldsymbol{\phi}$ is a distribution over words for each cluster, so it is optimized under the constraint that $\sum_{j=1}^V \phi_{z,j} = 1$. To incorporate this

constraint, we introduce a set of Lagrange multipliers $\{\lambda_z\}_{z=1}^K$, and from the Lagrangian,

$$J_\phi = \sum_{i=1}^N \sum_{z=1}^K q^{(i)}(z) \left(\log p(\mathbf{x}^{(i)} | z; \boldsymbol{\phi}) + \log p(z; \mu) - \log q^{(i)}(z) \right) + \sum_{z=1}^K \lambda_z \left(1 - \sum_{j=1}^V \phi_{z,j} \right). \quad [5.19]$$

2725 The term $\log p(\mathbf{x}^{(i)} | z; \boldsymbol{\phi})$ is the conditional log-likelihood for the multinomial, which
2726 expands to,

$$\log p(\mathbf{x}^{(i)} | z, \boldsymbol{\phi}) = C + \sum_{j=1}^V x_j \log \phi_{z,j}, \quad [5.20]$$

2727 where C is a constant with respect to $\boldsymbol{\phi}$ — see Equation 2.12 in § 2.2 for more discussion
2728 of this probability function.

Setting the derivative of J_ϕ equal to zero,

$$\frac{\partial J_\phi}{\partial \phi_{z,j}} = \sum_{i=1}^N q^{(i)}(z) \times \frac{x_j^{(i)}}{\phi_{z,j}} - \lambda_z \quad [5.21]$$

$$\phi_{z,j} \propto \sum_{i=1}^N q^{(i)}(z) \times x_j^{(i)}. \quad [5.22]$$

Because ϕ_z is constrained to be a probability distribution, the exact solution is computed as,

$$\phi_{z,j} = \frac{\sum_{i=1}^N q^{(i)}(z) \times x_j^{(i)}}{\sum_{j'=1}^V \sum_{i=1}^N q^{(i)}(z) \times x_{j'}^{(i)}} = \frac{E_q [\text{count}(z, j)]}{\sum_{j'=1}^V E_q [\text{count}(z, j')]} \quad [5.23]$$

2729 where the counter $j \in \{1, 2, \dots, V\}$ indexes over base features, such as words.

2730 This update sets ϕ_z equal to the relative frequency estimate of the *expected counts* under
2731 the distribution q . As in supervised Naïve Bayes, we can smooth these counts by adding
2732 a constant α . The update for μ is similar: $\mu_z \propto \sum_{i=1}^N q^{(i)}(z) = E_q [\text{count}(z)]$, which is the
2733 expected frequency of cluster z . These probabilities can also be smoothed. In sum, the
2734 M-step is just like Naïve Bayes, but with expected counts rather than observed counts.

2735 The multinomial likelihood $p(\mathbf{x} | z)$ can be replaced with other probability distribu-
2736 tions: for example, for continuous observations, a Gaussian distribution can be used. In
2737 some cases, there is no closed-form update to the parameters of the likelihood. One ap-
2738 proach is to run gradient-based optimization at each M-step; another is to simply take a
2739 single step along the gradient step and then return to the E-step (Berg-Kirkpatrick et al.,
2740 2010).

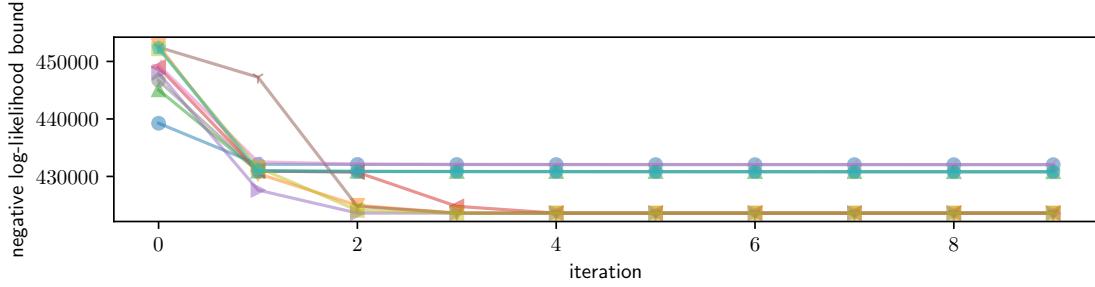


Figure 5.2: Sensitivity of expectation-maximization to initialization. Each line shows the progress of optimization from a different random initialization.

2741 5.1.3 EM as an optimization algorithm

2742 Algorithms that update a global objective by alternating between updates to subsets of the
 2743 parameters are called **coordinate ascent** algorithms. The objective J (the lower bound on
 2744 the marginal likelihood of the data) is separately convex in q and (μ, ϕ) , but it is not jointly
 2745 convex in all terms; this condition is known as **biconvexity**. Each step of the expectation-
 2746 maximization algorithm is guaranteed not to decrease the lower bound J , which means
 2747 that EM will converge towards a solution at which no nearby points yield further im-
 2748 provements. This solution is a **local optimum** — it is as good or better than any of its
 2749 immediate neighbors, but is *not* guaranteed to be optimal among all possible configura-
 2750 tions of (q, μ, ϕ) .

2751 The fact that there is no guarantee of global optimality means that initialization is
 2752 important: where you start can determine where you finish. To illustrate this point,
 2753 Figure 5.2 shows the objective function for EM with ten different random initializations:
 2754 while the objective function improves monotonically in each run, it converges to several
 2755 different values.¹ For the convex objectives that we encountered in chapter 2, it was not
 2756 necessary to worry about initialization, because gradient-based optimization guaranteed
 2757 to reach the global minimum. But in expectation-maximization — as in the deep neural
 2758 networks from chapter 3 — initialization matters.

2759 In **hard EM**, each $q^{(i)}$ distribution assigns probability of 1 to a single label $\hat{z}^{(i)}$, and zero
 2760 probability to all others (Neal and Hinton, 1998). This is similar in spirit to K -means clus-
 2761 tering, and can outperform standard EM in some cases (Spitkovsky et al., 2010). Another
 2762 variant of expectation-maximization incorporates stochastic gradient descent (SGD): after
 2763 performing a local E-step at each instance $x^{(i)}$, we immediately make a gradient update
 2764 to the parameters (μ, ϕ) . This algorithm has been called **incremental expectation maxi-**
 2765 **mization** (Neal and Hinton, 1998) and **online expectation maximization** (Sato and Ishii,

¹The figure shows the upper bound on the *negative* log-likelihood, because optimization is typically framed as minimization rather than maximization.

2766 2000; Cappé and Moulines, 2009), and is especially useful when there is no closed-form
 2767 optimum for the likelihood $p(\mathbf{x} | z)$, and in online settings where new data is constantly
 2768 streamed in (see Liang and Klein, 2009, for a comparison for online EM variants).

2769 **5.1.4 How many clusters?**

2770 So far, we have assumed that the number of clusters K is given. In some cases, this as-
 2771 sumption is valid. For example, a lexical semantic resource like WORDNET might define
 2772 the number of senses for a word. In other cases, the number of clusters could be a parame-
 2773 ter for the user to tune: some readers want a coarse-grained clustering of news stories into
 2774 three or four clusters, while others want a fine-grained clustering into twenty or more. But
 2775 many times there is little extrinsic guidance for how to choose K .

2776 One solution is to choose the number of clusters to maximize a metric of clustering
 2777 quality. The other parameters μ and ϕ are chosen to maximize the log-likelihood bound
 2778 J , so this might seem a potential candidate for tuning K . However, J will never decrease
 2779 with K : if it is possible to obtain a bound of J_K with K clusters, then it is always possible
 2780 to do at least as well with $K + 1$ clusters, by simply ignoring the additional cluster and
 2781 setting its probability to zero in q and μ . It is therefore necessary to introduce a penalty
 2782 for model complexity, so that fewer clusters are preferred. For example, the Akaike Infor-
 2783 mation Crition (AIC; Akaike, 1974) is the linear combination of the number of parameters
 2784 and the log-likelihood,

$$\text{AIC} = 2M - 2J, \quad [5.24]$$

2785 where M is the number of parameters. In an expectation-maximization clustering algo-
 2786 rithm, $M = K \times V + K$. Since the number of parameters increases with the number of
 2787 clusters K , the AIC may prefer more parsimonious models, even if they do not fit the data
 2788 quite as well.

2789 Another choice is to maximize the **predictive likelihood** on heldout data. This data
 2790 is not used to estimate the model parameters ϕ and μ , and so it is not the case that the
 2791 likelihood on this data is guaranteed to increase with K . Figure 5.3 shows the negative
 2792 log-likelihood on training and heldout data, as well as the AIC.

2793 ***Bayesian nonparametrics** An alternative approach is to treat the number of clusters
 2794 as another latent variable. This requires statistical inference over a set of models with a
 2795 variable number of clusters. This is not possible within the framework of expecta-
 2796 tion maximization, but there are several alternative inference procedures which can be ap-
 2797 plied, including **Markov Chain Monte Carlo (MCMC)**, which is briefly discussed in
 2798 § 5.5 (for more details, see Chapter 25 of Murphy, 2012). Bayesian nonparametrics have
 2799 been applied to the problem of unsupervised word sense induction, learning not only the
 2800 word senses but also the number of senses per word (Reisinger and Mooney, 2010).

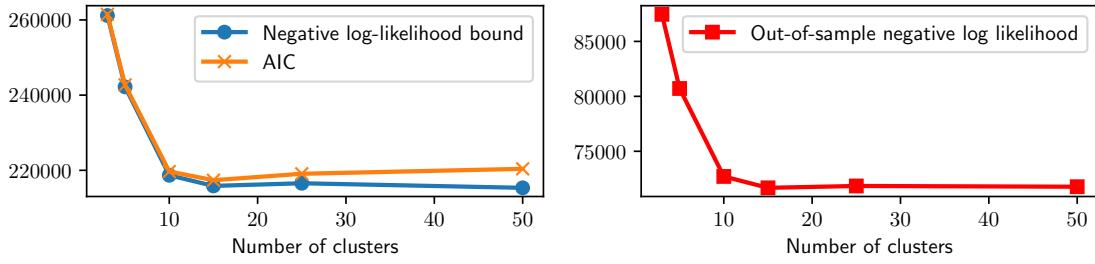


Figure 5.3: The negative log-likelihood and AIC for several runs of expectation-maximization, on synthetic data. Although the data was generated from a model with $K = 10$, the optimal number of clusters is $\hat{K} = 15$, according to AIC and the heldout log-likelihood. The training set log-likelihood continues to improve as K increases.

2801 5.2 Applications of expectation-maximization

2802 EM is not really an “algorithm” like, say, quicksort. Rather, it is a framework for learning
2803 with missing data. The recipe for using EM on a problem of interest is:

- 2804 • Introduce latent variables z , such that it is easy to write the probability $P(x, z)$. It
2805 should also be easy to estimate the associated parameters, given knowledge of z .
- 2806 • Derive the E-step updates for $q(z)$, which is typically factored as $q(z) = \prod_{i=1}^N q_{z^{(i)}}(z^{(i)})$,
2807 where i is an index over instances.
- 2808 • The M-step updates typically correspond to the soft version of a probabilistic super-
2809 vised learning algorithm, like Naïve Bayes.

2810 This section discusses a few of the many applications of this general framework.

2811 5.2.1 Word sense induction

2812 The chapter began by considering the problem of word sense disambiguation when the
2813 senses are not known in advance. Expectation-maximization can be applied to this prob-
2814 lem by treating each cluster as a word sense. Each instance represents the use of an
2815 ambiguous word, and $x^{(i)}$ is a vector of counts for the other words that appear nearby:
2816 Schütze (1998) uses all words within a 50-word window. The probability $p(x^{(i)} | z)$ can be
2817 set to the multinomial distribution, as in Naïve Bayes. The EM algorithm can be applied
2818 directly to this data, yielding clusters that (hopefully) correspond to the word senses.

Better performance can be obtained by first applying **singular value decomposition** (SVD) to the matrix of context-counts $C_{ij} = \text{count}(i, j)$, where $\text{count}(i, j)$ is the count of word j in the context of instance i . **Truncated** singular value decomposition approximates

the matrix \mathbf{C} as a product of three matrices, $\mathbf{U}, \mathbf{S}, \mathbf{V}$, under the constraint that \mathbf{U} and \mathbf{V} are orthonormal, and \mathbf{S} is diagonal:

$$\begin{aligned} & \min_{\mathbf{U}, \mathbf{S}, \mathbf{V}} \|\mathbf{C} - \mathbf{USV}^\top\|_F \\ & \text{s.t. } \mathbf{U} \in \mathbb{R}^{V \times K}, \mathbf{UU}^\top = \mathbb{I} \\ & \quad \mathbf{S} = \text{Diag}(s_1, s_2, \dots, s_K) \\ & \quad \mathbf{V}^\top \in \mathbb{R}^{N_p \times K}, \mathbf{VV}^\top = \mathbb{I}, \end{aligned} \quad [5.25]$$

where $\|\cdot\|_F$ is the **Frobenius norm**, $\|X\|_F = \sqrt{\sum_{i,j} X_{i,j}^2}$. The matrix \mathbf{U} contains the left singular vectors of \mathbf{C} , and the rows of this matrix can be used as low-dimensional representations of the count vectors \mathbf{c}_i . EM clustering can be made more robust by setting the instance descriptions $\mathbf{x}^{(i)}$ equal to these rows, rather than using raw counts (Schütze, 1998). However, because the instances are now dense vectors of continuous numbers, the probability $p(\mathbf{x}^{(i)} | z)$ must be defined as a multivariate Gaussian distribution.

In truncated singular value decomposition, the hyperparameter K is the truncation limit: when K is equal to the rank of \mathbf{C} , the norm of the difference between the original matrix \mathbf{C} and its reconstruction \mathbf{USV}^\top will be zero. Lower values of K increase the reconstruction error, but yield vector representations that are smaller and easier to learn from. Singular value decomposition is discussed in more detail in chapter 14.

5.2.2 Semi-supervised learning

Expectation-maximization can also be applied to the problem of **semi-supervised learning**: learning from both labeled and unlabeled data in a single model. Semi-supervised learning makes use of annotated examples, ensuring that each label y corresponds to the desired concept. By adding unlabeled examples, it is possible cover a greater fraction of the features than would appear in labeled data alone. Other methods for semi-supervised learning are discussed in § 5.3, but for now, let's approach the problem within the framework of expectation-maximization (Nigam et al., 2000).

Suppose we have labeled data $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{N_\ell}$, and unlabeled data $\{\mathbf{x}^{(i)}\}_{i=N_\ell+1}^{N_\ell+N_u}$, where N_ℓ is the number of labeled instances and N_u is the number of unlabeled instances. We can learn from the combined data by maximizing a lower bound on the joint log-likelihood,

$$\mathcal{L} = \sum_{i=1}^{N_\ell} \log p(\mathbf{x}^{(i)}, y^{(i)}; \boldsymbol{\mu}, \boldsymbol{\phi}) + \sum_{j=N_\ell+1}^{N_\ell+N_u} \log p(\mathbf{x}^{(j)}; \boldsymbol{\mu}, \boldsymbol{\phi}) \quad [5.26]$$

$$= \sum_{i=1}^{N_\ell} \left(\log p(\mathbf{x}^{(i)} | y^{(i)}; \boldsymbol{\phi}) + \log p(y^{(i)}; \boldsymbol{\mu}) \right) + \sum_{j=N_\ell+1}^{N_\ell+N_u} \log \sum_{y=1}^K p(\mathbf{x}^{(j)}, y; \boldsymbol{\mu}, \boldsymbol{\phi}). \quad [5.27]$$

Algorithm 9 Generative process for the Naïve Bayes classifier with hidden components

for Instance $i \in \{1, 2, \dots, N\}$ **do**:

Draw the label $y^{(i)} \sim \text{Categorical}(\mu)$;

Draw the component $z^{(i)} \sim \text{Categorical}(\beta_{y^{(i)}})$;

Draw the word counts $x^{(i)} | y^{(i)}, z^{(i)} \sim \text{Multinomial}(\phi_{z^{(i)}})$.

2838 The left sum is identical to the objective in Naïve Bayes; the right sum is the marginal log-
 2839 likelihood for expectation-maximization clustering, from Equation 5.5. We can construct a
 2840 lower bound on this log-likelihood by introducing distributions $q^{(j)}$ for all $j \in \{N_\ell + 1, \dots, N_\ell + N_u\}$.
 2841 The E-step updates these distributions; the M-step updates the parameters ϕ and μ , us-
 2842 ing the expected counts from the unlabeled data and the observed counts from the labeled
 2843 data.

2844 A critical issue in semi-supervised learning is how to balance the impact of the labeled
 2845 and unlabeled data on the classifier weights, especially when the unlabeled data is much
 2846 larger than the labeled dataset. The risk is that the unlabeled data will dominate, caus-
 2847 ing the parameters to drift towards a “natural clustering” of the instances — which may
 2848 not correspond to a good classifier for the labeled data. One solution is to heuristically
 2849 reweight the two components of Equation 5.26, tuning the weight of the two components
 2850 on a heldout development set (Nigam et al., 2000).

2851 **5.2.3 Multi-component modeling**

2852 As a final application, let’s return to fully supervised classification. A classic dataset for
 2853 text classification is 20 newsgroups, which contains posts to a set of online forums, called
 2854 newsgroups. One of the newsgroups is `comp.sys.mac.hardware`, which discusses Ap-
 2855 ple computing hardware. Suppose that within this newsgroup there are two kinds of
 2856 posts: reviews of new hardware, and question-answer posts about hardware problems.
 2857 The language in these *components* of the `mac.hardware` class might have little in com-
 2858 mon; if so, it would be better to model these components separately, rather than treating
 2859 their union as a single class. However, the component responsible for each instance is not
 2860 directly observed.

2861 Recall that Naïve Bayes is based on a generative process, which provides a stochastic
 2862 explanation for the observed data. In Naïve Bayes, each label is drawn from a categorical
 2863 distribution with parameter μ , and each vector of word counts is drawn from a multi-
 2864 nomial distribution with parameter ϕ_y . For multi-component modeling, we envision a
 2865 slightly different generative process, incorporating both the observed label $y^{(i)}$ and the
 2866 latent component $z^{(i)}$. This generative process is shown in Algorithm 9. A new parameter
 2867 $\beta_{y^{(i)}}$ defines the distribution of components, conditioned on the label $y^{(i)}$. The component,
 2868 and not the class label, then parametrizes the distribution over words.

-
- (5.1) ☺ Villeneuve a bel et bien **réussi** son pari de changer de perspectives tout en assurant une cohérence à la franchise.²
- (5.2) ☺ Il est également trop **long** et bancal dans sa narration, tiède dans ses intentions, et tirailé entre deux personnages et directions qui ne parviennent pas à coexister en harmonie.³
- (5.3) Denis Villeneuve a **réussi** une suite **parfaitemment** maîtrisée⁴
- (5.4) **Long, bavard**, hyper design, à peine agité (le comble de l'action : une bagarre dans la flotte), métaphysique et, surtout, ennuyeux jusqu'à la catalepsie.⁵
- (5.5) Une suite d'une écrasante puissance, mêlant **parfaitemment** le contemplatif au narratif.⁶
- (5.6) Le film impitoyablement **bavard** finit quand même par se taire quand se lève l'espèce de bouquet final où semble se déchaîner, comme en libre parcours de poulets décapiés, l'armée des graphistes numériques griffant nerveusement la palette graphique entre agonie et orgasme.⁷

Table 5.1: Labeled and unlabeled reviews of the films *Blade Runner 2049* and *Transformers: The Last Knight*.

The labeled data includes $(\mathbf{x}^{(i)}, y^{(i)})$, but not $z^{(i)}$, so this is another case of missing data. Again, we sum over the missing data, applying Jensen's inequality to as to obtain a lower bound on the log-likelihood,

$$\log p(\mathbf{x}^{(i)}, y^{(i)}) = \log \sum_{z=1}^{K_z} p(\mathbf{x}^{(i)}, y^{(i)}, z; \boldsymbol{\mu}, \boldsymbol{\phi}, \boldsymbol{\beta}) \quad [5.28]$$

$$\geq \log p(y^{(i)}; \boldsymbol{\mu}) + E_{q_{Z|Y}^{(i)}} [\log p(\mathbf{x}^{(i)} | z; \boldsymbol{\phi}) + \log p(z | y^{(i)}; \boldsymbol{\beta}) - \log q^{(i)}(z)]. \quad [5.29]$$

We are now ready to apply expectation-maximization. As usual, the E-step updates the distribution over the missing data, $q_{Z|Y}^{(i)}$. The M-step updates the parameters,

$$\beta_{y,z} = \frac{E_q [\text{count}(y, z)]}{\sum_{z'=1}^{K_z} E_q [\text{count}(y, z')]} \quad [5.30]$$

$$\phi_{z,j} = \frac{E_q [\text{count}(z, j)]}{\sum_{j'=1}^V E_q [\text{count}(z, j')]} \quad [5.31]$$

2869 5.3 Semi-supervised learning

2870 In semi-supervised learning, the learner makes use of both labeled and unlabeled data.
 2871 To see how this could help, suppose you want to do sentiment analysis in French. In Ta-

2872 ble 5.1, there are two labeled examples, one positive and one negative. From this data, a
 2873 learner could conclude that *réussi* is positive and *long* is negative. This isn't much! How-
 2874 ever, we can propagate this information to the unlabeled data, and potentially learn more.

- 2875 • If we are confident that *réussi* is positive, then we might guess that (5.3) is also posi-
 2876 tive.
- 2877 • That suggests that *parfaitement* is also positive.
- 2878 • We can then propagate this information to (5.5), and learn from this words in this
 2879 example.
- 2880 • Similarly, we can propagate from the labeled data to (5.4), which we guess to be
 2881 negative because it shares the word *long*. This suggests that *bavard* is also negative,
 2882 which we propagate to (5.6).

2883 Instances (5.3) and (5.4) were "similar" to the labeled examples for positivity and negativ-
 2884 ity, respectively. By using these instances to expand the models for each class, it became
 2885 possible to correctly label instances (5.5) and (5.6), which didn't share any important fea-
 2886 tures with the original labeled data. This requires a key assumption: that similar instances
 2887 will have similar labels.

2888 In § 5.2.2, we discussed how expectation-maximization can be applied to semi-supervised
 2889 learning. Using the labeled data, the initial parameters ϕ would assign a high weight for
 2890 *réussi* in the positive class, and a high weight for *long* in the negative class. These weights
 2891 helped to shape the distributions q for instances (5.3) and (5.4) in the E-step. In the next
 2892 iteration of the M-step, the parameters ϕ are updated with counts from these instances,
 2893 making it possible to correctly label the instances (5.5) and (5.6).

2894 However, expectation-maximization has an important disadvantage: it requires using
 2895 a generative classification model, which restricts the features that can be used for clas-
 2896 sification. In this section, we explore non-probabilistic approaches, which impose fewer
 2897 restrictions on the classification model.

2898 5.3.1 Multi-view learning

2899 EM semi-supervised learning can be viewed as **self-training**: the labeled data guides the
 2900 initial estimates of the classification parameters; these parameters are used to compute
 2901 a label distribution over the unlabeled instances, $q^{(i)}$; the label distributions are used to
 2902 update the parameters. The risk is that self-training drifts away from the original labeled
 2903 data. This problem can be ameliorated by **multi-view learning**. Here we take the as-
 2904 sumption that the features can be decomposed into multiple "views", each of which is
 2905 conditionally independent, given the label. For example, consider the problem of classi-
 2906 fying a name as a person or location: one view is the name itself; another is the context in
 2907 which it appears. This situation is illustrated in Table 5.2.

	$\mathbf{x}^{(1)}$	$\mathbf{x}^{(2)}$	y
1.	Peachtree Street	located on	LOC
2.	Dr. Walker	said	PER
3.	Zanzibar	located in	? → LOC
4.	Zanzibar	flew to	? → LOC
5.	Dr. Robert	recommended	? → PER
6.	Oprah	recommended	? → PER

Table 5.2: Example of multiview learning for named entity classification

2908 **Co-training** is an iterative multi-view learning algorithm, in which there are separate
 2909 classifiers for each view (Blum and Mitchell, 1998). At each iteration of the algorithm, each
 2910 classifier predicts labels for a subset of the unlabeled instances, using only the features
 2911 available in its view. These predictions are then used as ground truth to train the classifiers
 2912 associated with the other views. In the example shown in Table 5.2, the classifier on $\mathbf{x}^{(1)}$
 2913 might correctly label instance #5 as a person, because of the feature *Dr*; this instance would
 2914 then serve as training data for the classifier on $\mathbf{x}^{(2)}$, which would then be able to correctly
 2915 label instance #6, thanks to the feature *recommended*. If the views are truly independent,
 2916 this procedure is robust to drift. Furthermore, it imposes no restrictions on the classifiers
 2917 that can be used for each view.

2918 Word-sense disambiguation is particularly suited to multi-view learning, thanks to the
 2919 heuristic of “one sense per discourse”: if a polysemous word is used more than once in
 2920 a given text or conversation, all usages refer to the same sense (Gale et al., 1992). This
 2921 motivates a multi-view learning approach, in which one view corresponds to the local
 2922 context (the surrounding words), and another view corresponds to the global context at
 2923 the document level (Yarowsky, 1995). The local context view is first trained on a small
 2924 seed dataset. We then identify its most confident predictions on unlabeled instances. The
 2925 global context view is then used to extend these confident predictions to other instances
 2926 within the same documents. These new instances are added to the training data to the
 2927 local context classifier, which is retrained and then applied to the remaining unlabeled
 2928 data.

2929 5.3.2 Graph-based algorithms

2930 Another family of approaches to semi-supervised learning begins by constructing a graph,
 2931 in which pairs of instances are linked with symmetric weights $\omega_{i,j}$, e.g.,

$$\omega_{i,j} = \exp(-\alpha \times \|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\|^2). \quad [5.32]$$

2932 The goal is to use this weighted graph to propagate labels from a small set of labeled
 2933 instances to larger set of unlabeled instances.

2934 In **label propagation**, this is done through a series of matrix operations (Zhu et al.,
 2935 2003). Let \mathbf{Q} be a matrix of size $N \times K$, in which each row $\mathbf{q}^{(i)}$ describes the labeling
 2936 of instance i . When ground truth labels are available, then $\mathbf{q}^{(i)}$ is an indicator vector,
 2937 with $q_{y^{(i)}}^{(i)} = 1$ and $q_{y' \neq y^{(i)}}^{(i)} = 0$. Let us refer to the submatrix of rows containing labeled
 2938 instances as \mathbf{Q}_L , and the remaining rows as \mathbf{Q}_U . The rows of \mathbf{Q}_U are initialized to assign
 2939 equal probabilities to all labels, $q_{i,k} = \frac{1}{K}$.

2940 Now, let $T_{i,j}$ represent the “transition” probability of moving from node j to node i ,

$$T_{i,j} \triangleq \Pr(j \rightarrow i) = \frac{\omega_{i,j}}{\sum_{k=1}^N \omega_{k,j}}. \quad [5.33]$$

We compute values of $T_{i,j}$ for all instances j and all *unlabeled* instances i , forming a matrix
 of size $N_U \times N$. If the dataset is large, this matrix may be expensive to store and manip-
 ulate; a solution is to sparsify it, by keeping only the κ largest values in each row, and
 setting all other values to zero. We can then “propagate” the label distributions to the
 unlabeled instances,

$$\tilde{\mathbf{Q}}_U \leftarrow \mathbf{T}\mathbf{Q} \quad [5.34]$$

$$\mathbf{s} \leftarrow \tilde{\mathbf{Q}}_U \mathbf{1} \quad [5.35]$$

$$\mathbf{Q}_U \leftarrow \text{Diag}(\mathbf{s})^{-1} \tilde{\mathbf{Q}}_U. \quad [5.36]$$

2941 The expression $\tilde{\mathbf{Q}}_U \mathbf{1}$ indicates multiplication of $\tilde{\mathbf{Q}}_U$ by a column vector of ones, which is
 2942 equivalent to computing the sum of each row of $\tilde{\mathbf{Q}}_U$. The matrix $\text{Diag}(\mathbf{s})$ is a diagonal
 2943 matrix with the elements of \mathbf{s} on the diagonals. The product $\text{Diag}(\mathbf{s})^{-1} \tilde{\mathbf{Q}}_U$ has the effect
 2944 of normalizing the rows of $\tilde{\mathbf{Q}}_U$, so that each row of \mathbf{Q}_U is a probability distribution over
 2945 labels.

2946 5.4 Domain adaptation

2947 In many practical scenarios, the labeled data differs in some key respect from the data
 2948 to which the trained model is to be applied. A classic example is in consumer reviews:
 2949 we may have labeled reviews of movies (the source domain), but we want to predict the
 2950 reviews of appliances (the target domain). A similar issues arise with genre differences:
 2951 most linguistically-annotated data is news text, but application domains range from social
 2952 media to electronic health records. In general, there may be several source and target
 2953 domains, each with their own properties; however, for simplicity, this discussion will
 2954 focus mainly on the case of a single source and target domain.

2955 The simplest approach is “direct transfer”: train a classifier on the source domain, and
 2956 apply it directly to the target domain. The accuracy of this approach depends on the extent
 2957 to which features are shared across domains. In review text, words like *outstanding* and

2958 *disappointing* will apply across both movies and appliances; but others, like *terrifying*, may
 2959 have meanings that are domain-specific. As a result, direct transfer performs poorly: for
 2960 example, an out-of-domain classifier (trained on book reviews) suffers twice the error rate
 2961 of an in-domain classifier on reviews of kitchen appliances (Blitzer et al., 2007). **Domain**
 2962 **adaptation** algorithms attempt to do better than direct transfer by learning from data in
 2963 both domains. There are two main families of domain adaptation algorithms, depending
 2964 on whether any labeled data is available in the target domain.

2965 5.4.1 Supervised domain adaptation

2966 In supervised domain adaptation, there is a small amount of labeled data in the target
 2967 domain, and a large amount of data in the source domain. The simplest approach would
 2968 be to ignore domain differences, and simply merge the training data from the source and
 2969 target domains. There are several other baseline approaches to dealing with this sce-
 2970 nario (Daumé III, 2007):

2971 **Interpolation.** Train a classifier for each domain, and combine their predictions, e.g.,

$$\hat{y} = \operatorname{argmax}_y \lambda_s \Psi_s(\mathbf{x}, y) + (1 - \lambda_s) \Psi_t(\mathbf{x}, y), \quad [5.37]$$

2972 where Ψ_s and Ψ_t are the scoring functions from the source and target domain clas-
 2973 sifiers respectively, and λ_s is the interpolation weight.

Prediction. Train a classifier on the source domain data, use its prediction as an additional
 feature in a classifier trained on the target domain data,

$$\hat{y}_s = \operatorname{argmax}_y \Psi_s(\mathbf{x}, y) \quad [5.38]$$

$$\hat{y}_t = \operatorname{argmax}_y \Psi_t([\mathbf{x}; \hat{y}_s], y). \quad [5.39]$$

2974 **Priors.** Train a classifier on the source domain data, and use its weights as a prior distri-
 2975 bution on the weights of the classifier for the target domain data. This is equivalent
 2976 to regularizing the target domain weights towards the weights of the source domain
 2977 classifier (Chelba and Acero, 2006),

$$\ell(\boldsymbol{\theta}_t) = \sum_{i=1}^N \ell^{(i)}(\mathbf{x}^{(i)}, y^{(i)}; \boldsymbol{\theta}_t) + \lambda \|\boldsymbol{\theta}_t - \boldsymbol{\theta}_s\|_2^2, \quad [5.40]$$

2978 where $\ell^{(i)}$ is the prediction loss on instance i , and λ is the regularization weight.

An effective and “frustratingly simple” alternative is EASYADAPT (Daumé III, 2007), which creates copies of each feature: one for each domain and one for the cross-domain setting. For example, a negative review of the film *Wonder Woman* begins, *As boring and flavorless as a three-day-old grilled cheese sandwich...*⁸ The resulting bag-of-words feature vector would be,

$$\begin{aligned} f(\mathbf{x}, y, d) = \{ & (boring, \odot, \text{MOVIE}) : 1, (boring, \odot, *) : 1, \\ & (flavorless, \odot, \text{MOVIE}) : 1, (flavorless, \odot, *) : 1, \\ & (three-day-old, \odot, \text{MOVIE}) : 1, (three-day-old, \odot, *) : 1, \\ & \dots \}, \end{aligned}$$

with $(boring, \odot, \text{MOVIE})$ indicating the word *boring* appearing in a negative labeled document in the MOVIE domain, and $(boring, \odot, *)$ indicating the same word in a negative labeled document in *any* domain. It is up to the learner to allocate weight between the domain-specific and cross-domain features: for words that facilitate prediction in both domains, the learner will use the cross-domain features; for words that are relevant only to a single domain, the domain-specific features will be used. Any discriminative classifier can be used with these augmented features.⁹

5.4.2 Unsupervised domain adaptation

In unsupervised domain adaptation, there is no labeled data in the target domain. Unsupervised domain adaptation algorithms cope with this problem by trying to make the data from the source and target domains as similar as possible. This is typically done by learning a **projection function**, which puts the source and target data in a shared space, in which a learner can generalize across domains. This projection is learned from data in both domains, and is applied to the base features — for example, the bag-of-words in text classification. The projected features can then be used both for training and for prediction.

Linear projection

In linear projection, the cross-domain representation is constructed by a matrix-vector product,

$$g(\mathbf{x}^{(i)}) = \mathbf{U}\mathbf{x}^{(i)}. \quad [5.41]$$

The projected vectors $g(\mathbf{x}^{(i)})$ can then be used as base features during both training (from the source domain) and prediction (on the target domain).

⁸<http://www.colesmithey.com/capsules/2017/06/wonder-woman.HTML>, accessed October 9, 2017.

⁹EASYADAPT can be explained as a hierarchical Bayesian model, in which the weights for each domain are drawn from a shared prior (Finkel and Manning, 2009).

2999 The projection matrix \mathbf{U} can be learned in a number of different ways, but many ap-
 3000 proaches focus on compressing and reconstructing the base features (Ando and Zhang,
 3001 2005). For example, we can define a set of **pivot features**, which are typically chosen be-
 3002 cause they appear in both domains: in the case of review documents, pivot features might
 3003 include evaluative adjectives like *outstanding* and *disappointing* (Blitzer et al., 2007). For
 3004 each pivot feature j , we define an auxiliary problem of predicting whether the feature is
 3005 present in each example, using the remaining base features. Let ϕ_j denote the weights of
 3006 this classifier, and us horizontally concatenate the weights for each of the N_p pivot features
 3007 into a matrix $\Phi = [\phi_1, \phi_2, \dots, \phi_{N_p}]$.

3008 We then perform truncated singular value decomposition on Φ , as described in § 5.2.1,
 3009 obtaining $\Phi \approx \mathbf{U}\mathbf{S}\mathbf{V}^\top$. The rows of the matrix \mathbf{U} summarize information about each base
 3010 feature: indeed, the truncated singular value decomposition identifies a low-dimension
 3011 basis for the weight matrix Φ , which in turn links base features to pivot features. Sup-
 3012 pose that a base feature *reliable* occurs only in the target domain of appliance reviews.
 3013 Nonetheless, it will have a positive weight towards some pivot features (e.g., *outstanding*,
 3014 *recommended*), and a negative weight towards others (e.g., *worthless*, *unpleasant*). A base
 3015 feature such as *watchable* might have the same associations with the pivot features, and
 3016 therefore, $\mathbf{u}_{\text{reliable}} \approx \mathbf{u}_{\text{watchable}}$. The matrix \mathbf{U} can thus project the base features into a
 3017 space in which this information is shared.

3018 Non-linear projection

3019 Non-linear transformations of the base features can be accomplished by implementing
 3020 the transformation function as a deep neural network, which is trained from an auxiliary
 3021 objective.

3022 **Denoising objectives** One possibility is to train a projection function to reconstruct a
 3023 corrupted version of the original input. The original input can be corrupted in various
 3024 ways: by the addition of random noise (Glorot et al., 2011; Chen et al., 2012), or by the
 3025 deletion of features (Chen et al., 2012; Yang and Eisenstein, 2015). Denoising objectives
 3026 share many properties of the linear projection method described above: they enable the
 3027 projection function to be trained on large amounts of unlabeled data from the target do-
 3028 main, and allow information to be shared across the feature space, thereby reducing sen-
 3029 sitivity to rare and domain-specific features.

3030 **Adversarial objectives** The ultimate goal is for the transformed representations $g(x^{(i)})$
 3031 to be domain-general. This can be made an explicit optimization criterion by comput-
 3032 ing the similarity of transformed instances both within and between domains (Tzeng
 3033 et al., 2015), or by formulating an auxiliary classification task, in which the domain it-
 3034 self is treated as a label (Ganin et al., 2016). This setting is **adversarial**, because we want

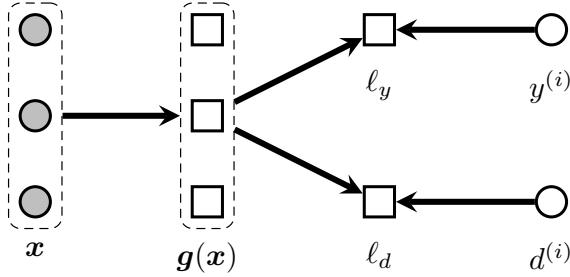


Figure 5.4: A schematic view of adversarial domain adaptation. The loss ℓ_y is computed only for instances from the source domain, where labels $y^{(i)}$ are available.

3035 to learn a representation that makes this classifier perform poorly. At the same time, we
 3036 want $g(x^{(i)})$ to enable accurate predictions of the labels $y^{(i)}$.

3037 To formalize this idea, let $d^{(i)}$ represent the domain of instance i , and let $\ell_d(g(x^{(i)}), d^{(i)}; \theta_d)$
 3038 represent the loss of a classifier (typically a deep neural network) trained to predict $d^{(i)}$
 3039 from the transformed representation $g(x^{(i)})$, using parameters θ_d . Analogously, let $\ell_y(g(x^{(i)}), y^{(i)}; \theta_y)$
 3040 represent the loss of a classifier trained to predict the label $y^{(i)}$ from $g(x^{(i)})$, using param-
 3041 eters θ_y . The transformation g can then be trained from two criteria: it should yield accu-
 3042 rate predictions of the labels $y^{(i)}$, while making *inaccurate* predictions of the domains $d^{(i)}$.
 3043 This can be formulated as a joint optimization problem,

$$\min_{\theta_g, \theta_y, \theta_d} \sum_{i=1}^{N_\ell + N_u} \ell_d(g(x^{(i)}; \theta_g), d^{(i)}; \theta_d) - \sum_{i=1}^{N_\ell} \ell_y(g(x^{(i)}; \theta_g), y^{(i)}; \theta_y), \quad [5.42]$$

3044 where N_ℓ is the number of labeled instances and N_u is the number of unlabeled instances,
 3045 with the labeled instances appearing first in the dataset. This setup is shown in Figure 5.4.
 3046 The loss can be optimized by stochastic gradient descent, jointly training the parameters
 3047 of the non-linear transformation θ_g , and the parameters of the prediction models θ_d and
 3048 θ_y .

3049 5.5 *Other approaches to learning with latent variables

3050 Expectation-maximization provides a general approach to learning with latent variables,
 3051 but it has limitations. One is the sensitivity to initialization; in practical applications,
 3052 considerable attention may need to be devoted to finding a good initialization. A second
 3053 issue is that EM tends to be easiest to apply in cases where the latent variables have a clear
 3054 decomposition (in the cases we have considered, they decompose across the instances).
 3055 For these reasons, it is worth briefly considering some alternatives to EM.

3056 **5.5.1 Sampling**

3057 In EM clustering, there is a distribution $\mathbf{q}^{(i)}$ for the missing data related to each instance.
 3058 The M-step consists of updating the parameters of this distribution. An alternative is to
 3059 draw samples of the latent variables. If the sampling distribution is designed correctly,
 3060 this procedure will eventually converge to drawing samples from the true posterior over
 3061 the missing data, $p(z^{(1:N_z)} | \mathbf{x}^{(1:N_x)})$. For example, in the case of clustering, the missing
 3062 data $\mathbf{z}^{(1:N_z)}$ is the set of cluster memberships, $\mathbf{y}^{(1:N)}$, so we draw samples from the pos-
 3063 terior distribution over clusterings of the data. If a single clustering is required, we can
 3064 select the one with the highest conditional likelihood, $\hat{\mathbf{z}} = \text{argmax}_{\mathbf{z}} p(\mathbf{z}^{(1:N_z)} | \mathbf{x}^{(1:N_x)})$.

This general family of algorithms is called **Markov Chain Monte Carlo (MCMC)**: “Monte Carlo” because it is based on a series of random draws; “Markov Chain” because the sampling procedure must be designed such that each sample depends only on the previous sample, and not on the entire sampling history. **Gibbs sampling** is an MCMC algorithm in which each latent variable is sampled from its posterior distribution,

$$z^{(n)} | \mathbf{x}, \mathbf{z}^{(-n)} \sim p(z^{(n)} | \mathbf{x}, \mathbf{z}^{(-n)}), \quad [5.43]$$

where $\mathbf{z}^{(-n)}$ indicates $\{\mathbf{z} \setminus z^{(n)}\}$, the set of all latent variables except for $z^{(n)}$. Repeatedly drawing samples over all latent variables constructs a Markov chain that is guaranteed to converge to a sequence of samples from $p(\mathbf{z}^{(1:N_z)} | \mathbf{x}^{(1:N_x)})$. In probabilistic clustering, the sampling distribution has the following form,

$$p(z^{(i)} | \mathbf{x}, \mathbf{z}^{(-i)}) = \frac{p(\mathbf{x}^{(i)} | z^{(i)}; \boldsymbol{\phi}) \times p(z^{(i)}; \boldsymbol{\mu})}{\sum_{z=1}^K p(\mathbf{x}^{(i)} | z; \boldsymbol{\phi}) \times p(z; \boldsymbol{\mu})} \quad [5.44]$$

$$\propto \text{Multinomial}(\mathbf{x}^{(i)}; \boldsymbol{\phi}_{z^{(i)}}) \times \boldsymbol{\mu}_{z^{(i)}}. \quad [5.45]$$

3065 In this case, the sampling distribution does not depend on the other instances: the poste-
 3066 rior distribution over each $z^{(i)}$ can be computed from $\mathbf{x}^{(i)}$ and the parameters given the
 3067 parameters $\boldsymbol{\phi}$ and $\boldsymbol{\mu}$.

3068 In sampling algorithms, there are several choices for how to deal with the parameters.
 3069 One possibility is to sample them too. To do this, we must add them to the generative
 3070 story, by introducing a prior distribution. For the multinomial and categorical parameters
 3071 in the EM clustering model, the **Dirichlet distribution** is a typical choice, since it defines
 3072 a probability on exactly the set of vectors that can be parameters: vectors that sum to one
 3073 and include only non-negative numbers.¹⁰

3074 To incorporate this prior, the generative model must be augmented to indicate that
 3075 each $\boldsymbol{\phi}_z \sim \text{Dirichlet}(\boldsymbol{\alpha}_\phi)$, and $\boldsymbol{\mu} \sim \text{Dirichlet}(\boldsymbol{\alpha}_\mu)$. The hyperparameters $\boldsymbol{\alpha}$ are typically set
 3076 to a constant vector $\boldsymbol{\alpha} = [\alpha, \alpha, \dots, \alpha]$. When α is large, the Dirichlet distribution tends to

¹⁰If $\sum_i^K \theta_i = 1$ and $\theta_i \geq 0$ for all i , then $\boldsymbol{\theta}$ is said to be on the $K - 1$ simplex. A Dirichlet distribution with

3077 generate vectors that are nearly uniform; when α is small, it tends to generate vectors that
 3078 assign most of their probability mass to a few entries. Given prior distributions over ϕ
 3079 and μ , we can now include them in Gibbs sampling, drawing values for these parameters
 3080 from posterior distributions that are conditioned on the other variables in the model.

3081 Unfortunately, sampling ϕ and μ usually leads to slow “mixing”, meaning that adja-
 3082 cent samples tend to be similar, so that a large number of samples is required to explore
 3083 the space of random variables. The reason is that the sampling distributions for the pa-
 3084 rameters are tightly constrained by the cluster memberships $y^{(i)}$, which in turn are tightly
 3085 constrained by the parameters. There are two solutions that are frequently employed:

- 3086 • **Empirical Bayesian** methods maintain ϕ and μ as parameters rather than latent
 3087 variables. They still employ sampling in the E-step of the EM algorithm, but they
 3088 update the parameters using expected counts that are computed from the samples
 3089 rather than from parametric distributions. This EM-MCMC hybrid is also known
 3090 as Monte Carlo Expectation Maximization (MCEM; Wei and Tanner, 1990), and is
 3091 well-suited for cases in which it is difficult to compute $q^{(i)}$ directly.
- 3092 • In **collapsed Gibbs sampling**, we analytically integrate ϕ and μ out of the model.
 3093 The cluster memberships $y^{(i)}$ are the only remaining latent variable; we sample them
 3094 from the compound distribution,

$$p(y^{(i)} | \mathbf{x}^{(1:N)}, \mathbf{y}^{(-i)}; \alpha_\phi, \alpha_\mu) = \int_{\phi, \mu} p(\phi, \mu | \mathbf{y}^{(-i)}, \mathbf{x}^{(1:N)}; \alpha_\phi, \alpha_\mu) p(y^{(i)} | \mathbf{x}^{(1:N)}, \mathbf{y}^{(-i)}, \phi, \mu) d\phi d\mu. \quad [5.48]$$

3095 For multinomial and Dirichlet distributions, this integral can be computed in closed
 3096 form.

3097 MCMC algorithms are guaranteed to converge to the true posterior distribution over
 3098 the latent variables, but there is no way to know how long this will take. In practice, the
 3099 rate of convergence depends on initialization, just as expectation-maximization depends
 3100 on initialization to avoid local optima. Thus, while Gibbs Sampling and other MCMC
 3101 algorithms provide a powerful and flexible array of techniques for statistical inference in
 3102 latent variable models, they are not a panacea for the problems experienced by EM.

parameter $\alpha \in \mathbb{R}_+^K$ has support over the $K - 1$ simplex,

$$p_{\text{Dirichlet}}(\boldsymbol{\theta} | \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^K \theta_i^{\alpha_i - 1} \quad [5.46]$$

$$B(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^K \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^K \alpha_i)}, \quad [5.47]$$

with $\Gamma(\cdot)$ indicating the gamma function, a generalization of the factorial function to non-negative reals.

3103 5.5.2 Spectral learning

Another approach to learning with latent variables is based on the **method of moments**, which makes it possible to avoid the problem of non-convex log-likelihood. Write $\bar{\mathbf{x}}^{(i)}$ for the normalized vector of word counts in document i , so that $\bar{\mathbf{x}}^{(i)} = \mathbf{x}^{(i)} / \sum_{j=1}^V x_j^{(i)}$. Then we can form a matrix of word-word co-occurrence probabilities,

$$\mathbf{C} = \sum_{i=1}^N \bar{\mathbf{x}}^{(i)} (\bar{\mathbf{x}}^{(i)})^\top. \quad [5.49]$$

The expected value of this matrix under $p(\mathbf{x} | \phi, \mu)$, as

$$E[\mathbf{C}] = \sum_{i=1}^N \sum_{k=1}^K \Pr(Z^{(i)} = k; \boldsymbol{\mu}) \phi_k \phi_k^\top \quad [5.50]$$

$$= \sum_k^K N \mu_k \phi_k \phi_k^\top \quad [5.51]$$

$$= \Phi \text{Diag}(N\mu) \Phi^\top, \quad [5.52]$$

where Φ is formed by horizontally concatenating $\phi_1 \dots \phi_K$, and $\text{Diag}(N\mu)$ indicates a diagonal matrix with values $N\mu_k$ at position (k, k) . Setting \mathbf{C} equal to its expectation gives,

$$\mathbf{C} = \Phi \text{Diag}(N\mu) \Phi^\top, \quad [5.53]$$

3104 which is similar to the eigendecomposition $\mathbf{C} = \mathbf{Q}\Lambda\mathbf{Q}^\top$. This suggests that simply by
 3105 finding the eigenvectors and eigenvalues of \mathbf{C} , we could obtain the parameters ϕ and μ ,
 3106 and this is what motivates the name **spectral learning**.

3107 While moment-matching and eigendecomposition are similar in form, they impose
 3108 different constraints on the solutions: eigendecomposition requires orthonormality, so
 3109 that $\mathbf{Q}\mathbf{Q}^\top = \mathbb{I}$; in estimating the parameters of a text clustering model, we require that μ
 3110 and the columns of Φ are probability vectors. Spectral learning algorithms must therefore
 3111 include a procedure for converting the solution into vectors that are non-negative and
 3112 sum to one. One approach is to replace eigendecomposition (or the related singular value
 3113 decomposition) with non-negative matrix factorization (Xu et al., 2003), which guarantees
 3114 that the solutions are non-negative (Arora et al., 2013).

3115 After obtaining the parameters ϕ and μ , the distribution over clusters can be com-
 3116 puted from Bayes' rule:

$$p(z^{(i)} | \mathbf{x}^{(i)}; \phi, \mu) \propto p(\mathbf{x}^{(i)} | z^{(i)}; \phi) \times p(z^{(i)}; \mu). \quad [5.54]$$

3117 Spectral learning yields provably good solutions without regard to initialization, and can
 3118 be quite fast in practice. However, it is more difficult to apply to a broad family of gener-
 3119 ative models than EM and Gibbs Sampling. For more on applying spectral learning across
 3120 a range of latent variable models, see Anandkumar et al. (2014).

3121 Additional resources

3122 There are a number of other learning paradigms that deviate from supervised learning.

- 3123 • **Active learning:** the learner selects unlabeled instances and requests annotations (Set-
 3124 tles, 2012).
- 3125 • **Multiple instance learning:** labels are applied to bags of instances, with a positive
 3126 label applied if at least one instance in the bag meets the criterion (Dietterich et al.,
 3127 1997; Maron and Lozano-Pérez, 1998).
- 3128 • **Constraint-driven learning:** supervision is provided in the form of explicit con-
 3129 straints on the learner (Chang et al., 2007; Ganchev et al., 2010).
- 3130 • **Distant supervision:** noisy labels are generated from an external resource (Mintz
 3131 et al., 2009, also see § 17.2.3).
- 3132 • **Multitask learning:** the learner induces a representation that can be used to solve
 3133 multiple classification tasks (Collobert et al., 2011).
- 3134 • **Transfer learning:** the learner must solve a classification task that differs from the
 3135 labeled data (Pan and Yang, 2010).

3136 Expectation-maximization was introduced by Dempster et al. (1977), and is discussed
 3137 in more detail by Murphy (2012). Like most machine learning treatments, Murphy focuses
 3138 on continuous observations and Gaussian likelihoods, rather than the discrete observa-
 3139 tions typically encountered in natural language processing. Murphy (2012) also includes
 3140 an excellent chapter on MCMC; for a textbook-length treatment, see Robert and Casella
 3141 (2013). For still more on Bayesian latent variable models, see Barber (2012), and for ap-
 3142 plications of Bayesian models to natural language processing, see Cohen (2016). Surveys
 3143 are available for semi-supervised learning (Zhu and Goldberg, 2009) and domain adapta-
 3144 tion (Søgaard, 2013), although both pre-date the current wave of interest in deep learning.

3145 Exercises

- 3146 1. Derive the expectation maximization update for the parameter μ in the EM cluster-
 3147 ing model.

- 3148 2. Derive the E-step and M-step updates for the following generative model. You may
 3149 assume that the labels $y^{(i)}$ are observed, but $z_m^{(i)}$ is not.

- 3150 • For each instance i ,

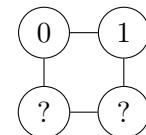
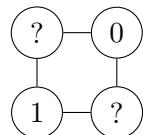
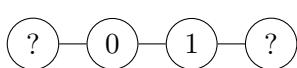
- 3151 – Draw label $y^{(i)} \sim \text{Categorical}(\mu)$
 3152 – For each token $m \in \{1, 2, \dots, M^{(i)}\}$

3153 * Draw $z_m^{(i)} \sim \text{Categorical}(\pi)$

3154 * If $z_m^{(i)} = 0$, draw the current token from a label-specific distribution,
 3155 $w_m^{(i)} \sim \phi_{y^{(i)}}$

3156 * If $z_m^{(i)} = 1$, draw the current token from a document-specific distribu-
 3157 tion, $w_m^{(i)} \sim \nu^{(i)}$

- 3158 3. Using the iterative updates in Equations 5.34-5.36, compute the outcome of the label
 3159 propagation algorithm for the following examples.



3160 The value inside the node indicates the label, $y^{(i)} \in \{0, 1\}$, with $y^{(i)} = ?$ for unlabeled
 3161 nodes. The presence of an edge between two nodes indicates $w_{i,j} = 1$, and the
 3162 absence of an edge indicates $w_{i,j} = 0$. For the third example, you need only compute
 3163 the first three iterations, and then you can guess at the solution in the limit.

- 3164 4. Use expectation-maximization clustering to train a word-sense induction system,
 3165 applied to the word *say*.

- 3166 • Import NLTK, run NLTK.DOWNLOAD() and select SEMCOR. Import SEMCOR
 3167 from NLTK.CORPUS.
- 3168 • The command SEMCOR.TAGGED_SENTENCES(TAG='SENSE') returns an itera-
 3169 tor over sense-tagged sentences in the corpus. Each sentence can be viewed
 3170 as an iterator over TREE objects. For TREE objects that are sense-annotated
 3171 words, you can access the annotation as TREE.LABEL(), and the word itself with
 3172 TREE.LEAVES(). So SEMCOR.TAGGED_SENTENCES(TAG='SENSE')[0][2].LABEL()
 3173 would return the sense annotation of the third word in the first sentence.
- 3174 • Extract all sentences containing the senses SAY.V.01 and SAY.V.02.
- 3175 • Build bag-of-words vectors $x^{(i)}$, containing the counts of other words in those
 3176 sentences, including all words that occur in at least two sentences.
- 3177 • Implement and run expectation-maximization clustering on the merged data.

- 3178 • Compute the frequency with which each cluster includes instances of SAY.V.01
 3179 and SAY.V.02.

3180 In the remaining exercises, you will try out some approaches for semisupervised learn-
 3181 ing and domain adaptation. You will need datasets in multiple domains. You can obtain
 3182 product reviews in multiple domains here: [https://www.cs.jhu.edu/~mdredze/](https://www.cs.jhu.edu/~mdredze/datasets/sentiment/processed_acl.tar.gz)
 3183 datasets/sentiment/processed_acl.tar.gz. Choose a source and target domain,
 3184 e.g. dvds and books, and divide the data for the target domain into training and test sets
 3185 of equal size.

3186 5. First, quantify the cost of cross-domain transfer.

- 3187 • Train a logistic regression classifier on the source domain training set, and eval-
 3188 uate it on the target domain test set.
 3189 • Train a logistic regression classifier on the target domain training set, and eval-
 3190 uate it on the target domain test set. This is the “direct transfer” baseline.

3191 Compute the difference in accuracy, which is a measure of the transfer loss across
 3192 domains.

3193 6. Next, apply the **label propagation** algorithm from § 5.3.2.

3194 As a baseline, using only 5% of the target domain training set, train a classifier, and
 3195 compute its accuracy on the target domain test set.

3196 Next, apply label propagation:

- 3197 • Compute the label matrix \mathbf{Q}_L for the labeled data (5% of the target domain
 3198 training set), with each row equal to an indicator vector for the label (positive
 3199 or negative).
 3200 • Iterate through the target domain instances, including both test and training
 3201 data. At each instance i , compute all w_{ij} , using Equation 5.32, with $\alpha = 0.01$.
 3202 Use these values to fill in column i of the transition matrix \mathbf{T} , setting all but the
 3203 ten largest values to zero for each column i . Be sure to normalize the column
 3204 so that the remaining values sum to one. You may need to use a sparse matrix
 3205 for this to fit into memory.
 3206 • Apply the iterative updates from Equations 5.34–5.36 to compute the outcome
 3207 of the label propagation algorithm for the unlabeled examples.

3208 Select the test set instances from \mathbf{Q}_U , and compute the accuracy of this method.
 3209 Compare with the supervised classifier trained only on the 5% sample of the target
 3210 domain training set.

- 3211 7. Using only 5% of the target domain training data (and all of the source domain train-
3212 ing data), implement one of the supervised domain adaptation baselines in § 5.4.1.
3213 See if this improves on the “direct transfer” baseline from the previous problem
- 3214 8. Implement EASYADAPT (§ 5.4.1), again using 5% of the target domain training data
3215 and all of the source domain data.
- 3216 9. Now try unsupervised domain adaptation, using the “linear projection” method
3217 described in § 5.4.2. Specifically:
- 3218 • Identify 500 pivot features as the words with the highest frequency in the (com-
3219 plete) training data for the source and target domains. Specifically, let x_i^d be the
3220 count of the word i in domain d : choose the 500 words with the largest values
3221 of $\min(x_i^{\text{source}}, x_i^{\text{target}})$.
- 3222 • Train a classifier to predict each pivot feature from the remaining words in the
3223 document.
- 3224 • Arrange the features of these classifiers into a matrix Φ , and perform truncated
3225 singular value decomposition, with $k = 20$
- 3226 • Train a classifier from the source domain data, using the combined features
3227 $\mathbf{x}^{(i)} \oplus \mathbf{U}^\top \mathbf{x}^{(i)}$ — these include the original bag-of-words features, plus the pro-
3228 jected features.
- 3229 • Apply this classifier to the target domain test set, and compute the accuracy.

3230

Part II

3231

Sequences and trees

3232 Chapter 6

3233 Language models

3234 In probabilistic classification, the problem is to compute the probability of a label, conditioned
3235 on the text. Let's now consider the inverse problem: computing the probability of
3236 text itself. Specifically, we will consider models that assign probability to a sequence of
3237 word tokens, $p(w_1, w_2, \dots, w_M)$, with $w_m \in \mathcal{V}$. The set \mathcal{V} is a discrete vocabulary,

$$\mathcal{V} = \{aardvark, abacus, \dots, zither\}. \quad [6.1]$$

3238 Why would you want to compute the probability of a word sequence? In many applications,
3239 the goal is to produce word sequences as output:

- 3240 • In **machine translation** (chapter 18), we convert from text in a source language to
3241 text in a target language.
- 3242 • In **speech recognition**, we convert from audio signal to text.
- 3243 • In **summarization** (§ 16.3.4; § 19.2), we convert from long texts into short texts.
- 3244 • In **dialogue systems** (§ 19.3), we convert from the user's input (and perhaps an
3245 external knowledge base) into a text response.

3246 In many of the systems for performing these tasks, there is a subcomponent that computes
3247 the probability of the output text. The purpose of this component is to generate
3248 texts that are more **fluent**. For example, suppose we want to translate a sentence from
3249 Spanish to English.

3250 (6.1) El cafe negro me gusta mucho.

3251 Here is a literal word-for-word translation (a **gloss**):

3252 (6.2) The coffee black me pleases much.

3253 A good language model of English will tell us that the probability of this translation is
 3254 low, in comparison with more grammatical alternatives,

$$p(\text{The coffee black me pleases much}) < p(\text{I love dark coffee}). \quad [6.2]$$

3255 How can we use this fact? Warren Weaver, one of the early leaders in machine trans-
 3256 lation, viewed it as a problem of breaking a secret code (Weaver, 1955):

3257 When I look at an article in Russian, I say: 'This is really written in English,
 3258 but it has been coded in some strange symbols. I will now proceed to decode.'

3259 This observation motivates a generative model (like Naïve Bayes):

- 3260 • The English sentence $w^{(e)}$ is generated from a **language model**, $p_e(w^{(e)})$.
- 3261 • The Spanish sentence $w^{(s)}$ is then generated from a **translation model**, $p_{s|e}(w^{(s)} | w^{(e)})$.

Given these two distributions, translation can be performed by Bayes' rule:

$$p_{e|s}(w^{(e)} | w^{(s)}) \propto p_{e,s}(w^{(e)}, w^{(s)}) \quad [6.3]$$

$$= p_{s|e}(w^{(s)} | w^{(e)}) \times p_e(w^{(e)}). \quad [6.4]$$

3262 This is sometimes called the **noisy channel model**, because it envisions English text
 3263 turning into Spanish by passing through a noisy channel, $p_{s|e}$. What is the advantage of
 3264 modeling translation this way, as opposed to modeling $p_{e|s}$ directly? The crucial point is
 3265 that the two distributions $p_{s|e}$ (the translation model) and p_e (the language model) can be
 3266 estimated from separate data. The translation model requires examples of correct trans-
 3267 lations, but the language model requires only text in English. Such monolingual data is
 3268 much more widely available. Furthermore, once estimated, the language model p_e can
 3269 be reused in any application that involves generating English text, including translation
 3270 from other languages.

3271 6.1 *N*-gram language models

A simple approach to computing the probability of a sequence of tokens is to use a **relative frequency estimate**. Consider the quote, attributed to Picasso, "*computers are useless, they can only give you answers.*" One way to estimate the probability of this sentence is,

$$\begin{aligned} p(\text{Computers are useless, they can only give you answers}) \\ = \frac{\text{count}(\text{Computers are useless, they can only give you answers})}{\text{count}(\text{all sentences ever spoken})} \end{aligned} \quad [6.5]$$

3272 This estimator is **unbiased**: in the theoretical limit of infinite data, the estimate will
 3273 be correct. But in practice, we are asking for accurate counts over an infinite number of
 3274 events, since sequences of words can be arbitrarily long. Even with an aggressive upper
 3275 bound of, say, $M = 20$ tokens in the sequence, the number of possible sequences is V^{20} ,
 3276 where $V = |\mathcal{V}|$. A small vocabulary for English would have $V = 10^5$, so there are 10^{100}
 3277 possible sequences. Clearly, this estimator is very data-hungry, and suffers from high vari-
 3278 ance: even grammatical sentences will have probability zero if they have not occurred in
 3279 the training data.¹ We therefore need to introduce bias to have a chance of making reli-
 3280 able estimates from finite training data. The language models that follow in this chapter
 3281 introduce bias in various ways.

We begin with n -gram language models, which compute the probability of a sequence as the product of probabilities of subsequences. The probability of a sequence $p(\mathbf{w}) = p(w_1, w_2, \dots, w_M)$ can be refactored using the chain rule (see § A.2):

$$p(\mathbf{w}) = p(w_1, w_2, \dots, w_M) \quad [6.6]$$

$$= p(w_1) \times p(w_2 | w_1) \times p(w_3 | w_2, w_1) \times \dots \times p(w_M | w_{M-1}, \dots, w_1) \quad [6.7]$$

Each element in the product is the probability of a word given all its predecessors. We can think of this as a *word prediction* task: given the context *Computers are*, we want to compute a probability over the next token. The relative frequency estimate of the probability of the word *useless* in this context is,

$$\begin{aligned} p(\text{useless} | \text{computers are}) &= \frac{\text{count}(\text{computers are useless})}{\sum_{x \in \mathcal{V}} \text{count}(\text{computers are } x)} \\ &= \frac{\text{count}(\text{computers are useless})}{\text{count}(\text{computers are})}. \end{aligned}$$

3282 We haven't made any approximations yet, and we could have just as well applied the
 3283 chain rule in reverse order,

$$p(\mathbf{w}) = p(w_M) \times p(w_{M-1} | w_M) \times \dots \times p(w_1 | w_2, \dots, w_M), \quad [6.8]$$

3284 or in any other order. But this means that we also haven't really made any progress:
 3285 to compute the conditional probability $p(w_M | w_{M-1}, w_{M-2}, \dots, w_1)$, we would need to
 3286 model V^{M-1} contexts. Such a distribution cannot be estimated from any realistic sample
 3287 of text.

To solve this problem, n -gram models make a crucial simplifying approximation: they condition on only the past $n - 1$ words.

$$p(w_m | w_{m-1} \dots w_1) \approx p(w_m | w_{m-1}, \dots, w_{m-n+1}) \quad [6.9]$$

¹Chomsky famously argued that this is evidence against the very concept of probabilistic language models: no such model could distinguish the grammatical sentence *colorless green ideas sleep furiously* from the ungrammatical permutation *furiously sleep ideas green colorless*.

This means that the probability of a sentence w can be approximated as

$$p(w_1, \dots, w_M) \approx \prod_{m=1}^M p(w_m | w_{m-1}, \dots, w_{m-n+1}) \quad [6.10]$$

To compute the probability of an entire sentence, it is convenient to pad the beginning and end with special symbols \square and \blacksquare . Then the bigram ($n = 2$) approximation to the probability of *I like black coffee* is:

$$p(I \text{ like black coffee}) = p(I | \square) \times p(\text{like} | I) \times p(\text{black} | \text{like}) \times p(\text{coffee} | \text{black}) \times p(\blacksquare | \text{coffee}). \quad [6.11]$$

3288 This model requires estimating and storing the probability of only V^n events, which is
 3289 exponential in the order of the n -gram, and not V^M , which is exponential in the length of
 3290 the sentence. The n -gram probabilities can be computed by relative frequency estimation,

$$p(w_m | w_{m-1}, w_{m-2}) = \frac{\text{count}(w_{m-2}, w_{m-1}, w_m)}{\sum_{w'} \text{count}(w_{m-2}, w_{m-1}, w')} \quad [6.12]$$

3291 The hyperparameter n controls the size of the context used in each conditional proba-
 3292 bility. If this is misspecified, the language model will perform poorly. Let's consider the
 3293 potential problems concretely.

3294 **When n is too small.** Consider the following sentences:

3295 (6.3) **Gorillas** always like to groom **their** friends.

3296 (6.4) The **computer** that's on the 3rd floor of our office building **crashed**.

3297 In each example, the words written in bold depend on each other: the likelihood
 3298 of *their* depends on knowing that *gorillas* is plural, and the likelihood of *crashed* de-
 3299 pends on knowing that the subject is a *computer*. If the n -grams are not big enough
 3300 to capture this context, then the resulting language model would offer probabili-
 3301 ties that are too low for these sentences, and too high for sentences that fail basic
 3302 linguistic tests like number agreement.

3303 **When n is too big.** In this case, it is hard to get good estimates of the n -gram parameters from
 3304 our dataset, because of data sparsity. To handle the *gorilla* example, it is necessary to
 3305 model 6-grams, which means accounting for V^6 events. Under a very small vocab-
 3306 ularly of $V = 10^4$, this means estimating the probability of 10^{24} distinct events.

These two problems point to another **bias-variance tradeoff** (see § 2.2.4). A small n -gram size introduces high bias, and a large n -gram size introduces high variance. We can even have both problems at the same time! Language is full of long-range dependencies that we cannot capture because n is too small; at the same time, language datasets are full of rare phenomena, whose probabilities we fail to estimate accurately because n is too large. One solution is to try to keep n large, while still making low-variance estimates of the underlying parameters. To do this, we will introduce a different sort of bias: **smoothing**.

6.2 Smoothing and discounting

Limited data is a persistent problem in estimating language models. In § 6.1, we presented n -grams as a partial solution. Bit sparse data can be a problem even for low-order n -grams; at the same time, many linguistic phenomena, like subject-verb agreement, cannot be incorporated into language models without high-order n -grams. It is therefore necessary to add additional inductive biases to n -gram language models. This section covers some of the most intuitive and common approaches, but there are many more (see Chen and Goodman, 1999).

6.2.1 Smoothing

A major concern in language modeling is to avoid the situation $p(w) = 0$, which could arise as a result of a single unseen n-gram. A similar problem arose in Naïve Bayes, and the solution was **smoothing**: adding imaginary “pseudo” counts. The same idea can be applied to n -gram language models, as shown here in the bigram case,

$$p_{\text{smooth}}(w_m \mid w_{m-1}) = \frac{\text{count}(w_{m-1}, w_m) + \alpha}{\sum_{w' \in \mathcal{V}} \text{count}(w_{m-1}, w') + V\alpha}. \quad [6.13]$$

This basic framework is called **Lidstone smoothing**, but special cases have other names:

- **Laplace smoothing** corresponds to the case $\alpha = 1$.
- **Jeffreys-Perks law** corresponds to the case $\alpha = 0.5$, which works well in practice and benefits from some theoretical justification (Manning and Schütze, 1999).

To ensure that the probabilities are properly normalized, anything that we add to the numerator (α) must also appear in the denominator ($V\alpha$). This idea is reflected in the concept of **effective counts**:

$$c_i^* = (c_i + \alpha) \frac{M}{M + V\alpha}, \quad [6.14]$$

	counts	unsmoothed probability	Lidstone smoothing, $\alpha = 0.1$		Discounting, $d = 0.1$	
			effective counts	smoothed probability	effective counts	smoothed probability
<i>impropriety</i>	8	0.4	7.826	0.391	7.9	0.395
<i>offense</i>	5	0.25	4.928	0.246	4.9	0.245
<i>damage</i>	4	0.2	3.961	0.198	3.9	0.195
<i>deficiencies</i>	2	0.1	2.029	0.101	1.9	0.095
<i>outbreak</i>	1	0.05	1.063	0.053	0.9	0.045
<i>infirmity</i>	0	0	0.097	0.005	0.25	0.013
<i>cephalopods</i>	0	0	0.097	0.005	0.25	0.013

Table 6.1: Example of Lidstone smoothing and absolute discounting in a bigram language model, for the context $(\text{alleged}, \cdot)$, for a toy corpus with a total of twenty counts over the seven words shown. Note that discounting decreases the probability for all but the unseen words, while Lidstone smoothing increases the effective counts and probabilities for *deficiencies* and *outbreak*.

where c_i is the count of event i , c_i^* is the effective count, and $M = \sum_{i=1}^V c_i$ is the total number of tokens in the dataset (w_1, w_2, \dots, w_M) . This term ensures that $\sum_{i=1}^V c_i^* = \sum_{i=1}^V c_i = M$. The **discount** for each n-gram is then computed as,

$$d_i = \frac{c_i^*}{c_i} = \frac{(c_i + \alpha)}{c_i} \frac{M}{(M + V\alpha)}.$$

3335 6.2.2 Discounting and backoff

3336 Discounting “borrows” probability mass from observed n -grams and redistributes it. In
 3337 Lidstone smoothing, the borrowing is done by increasing the denominator of the relative
 3338 frequency estimates. The borrowed probability mass is then redistributed by increasing
 3339 the numerator for all n -grams. Another approach would be to borrow the same amount
 3340 of probability mass from all observed n -grams, and redistribute it among only the unob-
 3341 served n -grams. This is called **absolute discounting**. For example, suppose we set an
 3342 absolute discount $d = 0.1$ in a bigram model, and then redistribute this probability mass
 3343 equally over the unseen words. The resulting probabilities are shown in Table 6.1.

Discounting reserves some probability mass from the observed data, and we need not redistribute this probability mass equally. Instead, we can **backoff** to a lower-order language model: if you have trigrams, use trigrams; if you don’t have trigrams, use bigrams; if you don’t even have bigrams, use unigrams. This is called **Katz backoff**. In the simple

case of backing off from bigrams to unigrams, the bigram probabilities are,

$$c^*(i, j) = c(i, j) - d \quad [6.15]$$

$$p_{\text{Katz}}(i | j) = \begin{cases} \frac{c^*(i, j)}{c(j)} & \text{if } c(i, j) > 0 \\ \alpha(j) \times \frac{p_{\text{unigram}}(i)}{\sum_{i': c(i', j)=0} p_{\text{unigram}}(i')} & \text{if } c(i, j) = 0. \end{cases} \quad [6.16]$$

3344 The term $\alpha(j)$ indicates the amount of probability mass that has been discounted for
 3345 context j . This probability mass is then divided across all the unseen events, $\{i' : c(i', j) =$
 3346 $0\}$, proportional to the unigram probability of each word i' . The discount parameter d can
 3347 be optimized to maximize performance (typically held-out log-likelihood) on a develop-
 3348 ment set.

3349 6.2.3 *Interpolation

3350 Backoff is one way to combine different order n -gram models. An alternative approach
 3351 is **interpolation**: setting the probability of a word in context to a weighted sum of its
 3352 probabilities across progressively shorter contexts.

Instead of choosing a single n for the size of the n -gram, we can take the weighted average across several n -gram probabilities. For example, for an interpolated trigram model,

$$\begin{aligned} p_{\text{Interpolation}}(w_m | w_{m-1}, w_{m-2}) &= \lambda_3 p_3^*(w_m | w_{m-1}, w_{m-2}) \\ &\quad + \lambda_2 p_2^*(w_m | w_{m-1}) \\ &\quad + \lambda_1 p_1^*(w_m). \end{aligned}$$

3353 In this equation, p_n^* is the unsmoothed empirical probability given by an n -gram lan-
 3354 guage model, and λ_n is the weight assigned to this model. To ensure that the interpolated
 3355 $p(w)$ is still a valid probability distribution, the values of λ must obey the constraint,
 3356 $\sum_{n=1}^{n_{\max}} \lambda_n = 1$. But how to find the specific values?

3357 An elegant solution is **expectation-maximization**. Recall from chapter 5 that we can
 3358 think about EM as learning with *missing data*: we just need to choose missing data such
 3359 that learning would be easy if it weren't missing. What's missing in this case? Think of
 3360 each word w_m as drawn from an n -gram of unknown size, $z_m \in \{1 \dots n_{\max}\}$. This z_m is
 3361 the missing data that we are looking for. Therefore, the application of EM to this problem
 3362 involves the following **generative model**:

3363 **for** Each token $w_m, m = 1, 2, \dots, M$ **do**:
 3364 draw the n -gram size $z_m \sim \text{Categorical}(\lambda)$;
 3365 draw $w_m \sim p_{z_m}^*(w_m | w_{m-1}, \dots, w_{m-z_m})$.

If the missing data $\{Z_m\}$ were known, then λ could be estimated as the relative frequency,

$$\lambda_z = \frac{\text{count}(Z_m = z)}{M} \quad [6.17]$$

$$\propto \sum_{m=1}^M \delta(Z_m = z). \quad [6.18]$$

But since we do not know the values of the latent variables Z_m , we impute a distribution q_m in the E-step, which represents the degree of belief that word token w_m was generated from a n -gram of order z_m ,

$$q_m(z) \triangleq \Pr(Z_m = z \mid \mathbf{w}_{1:m}; \lambda) \quad [6.19]$$

$$= \frac{p(w_m \mid \mathbf{w}_{1:m-1}, Z_m = z) \times p(z)}{\sum_{z'} p(w_m \mid \mathbf{w}_{1:m-1}, Z_m = z') \times p(z')} \quad [6.20]$$

$$\propto p_z^*(w_m \mid \mathbf{w}_{1:m-1}) \times \lambda_z. \quad [6.21]$$

In the M-step, λ is computed by summing the expected counts under q ,

$$\lambda_z \propto \sum_{m=1}^M q_m(z). \quad [6.22]$$

3367 A solution is obtained by iterating between updates to q and λ . The complete algorithm
 3368 is shown in Algorithm 10.

Algorithm 10 Expectation-maximization for interpolated language modeling

```

1: procedure ESTIMATE INTERPOLATED  $n$ -GRAM ( $\mathbf{w}_{1:M}, \{p_n^*\}_{n \in 1:n_{\max}}$ )
2:   for  $z \in \{1, 2, \dots, n_{\max}\}$  do ▷ Initialization
3:      $\lambda_z \leftarrow \frac{1}{n_{\max}}$ 
4:   repeat
5:     for  $m \in \{1, 2, \dots, M\}$  do ▷ E-step
6:       for  $z \in \{1, 2, \dots, n_{\max}\}$  do
7:          $q_m(z) \leftarrow p_z^*(w_m \mid \mathbf{w}_{1:m-1}) \times \lambda_z$ 
8:        $q_m \leftarrow \text{Normalize}(q_m)$ 
9:     for  $z \in \{1, 2, \dots, n_{\max}\}$  do ▷ M-step
10:       $\lambda_z \leftarrow \frac{1}{M} \sum_{m=1}^M q_m(z)$ 
11:    until tired
12:    return  $\lambda$ 
  
```

3369 **6.2.4 *Kneser-Ney smoothing**

3370 **Kneser-Ney smoothing** is based on absolute discounting, but it redistributes the result-
 3371 ing probability mass in a different way from Katz backoff. Empirical evidence points
 3372 to Kneser-Ney smoothing as the state-of-art for n -gram language modeling (Goodman,
 3373 2001). To motivate Kneser-Ney smoothing, consider the example: *I recently visited ..*
 3374 Which of the following is more likely: *Francisco* or *Duluth*?

3375 Now suppose that both bigrams *visited Duluth* and *visited Francisco* are unobserved in
 3376 the training data, and furthermore, that the unigram probability $p_1^*(\text{Francisco})$ is greater
 3377 than $p_1^*(\text{Duluth})$. Nonetheless we would still guess that $p(\text{visited Duluth}) > p(\text{visited Francisco})$,
 3378 because *Duluth* is a more “versatile” word: it can occur in many contexts, while *Francisco*
 3379 usually occurs in a single context, following the word *San*. This notion of versatility is the
 3380 key to Kneser-Ney smoothing.

Writing u for a context of undefined length, and $\text{count}(w, u)$ as the count of word w in
 context u , we define the Kneser-Ney bigram probability as

$$p_{KN}(w | u) = \begin{cases} \frac{\max(\text{count}(w, u) - d, 0)}{\text{count}(u)}, & \text{count}(w, u) > 0 \\ \alpha(u) \times p_{\text{continuation}}(w), & \text{otherwise} \end{cases} \quad [6.23]$$

$$p_{\text{continuation}}(w) = \frac{|u : \text{count}(w, u) > 0|}{\sum_{w' \in \mathcal{V}} |u' : \text{count}(w', u') > 0|}. \quad [6.24]$$

3381 Probability mass using absolute discounting d , which is taken from all unobserved
 3382 n -grams. The total amount of discounting in context u is $d \times |w : \text{count}(w, u) > 0|$, and
 3383 we divide this probability mass among the unseen n -grams. To account for versatility,
 3384 we define the *continuation probability* $p_{\text{continuation}}(w)$ as proportional to the number of ob-
 3385 served contexts in which w appears. The numerator of the continuation probability is the
 3386 number of contexts u in which w appears; the denominator normalizes the probability by
 3387 summing the same quantity over all words w' . The coefficient $\alpha(u)$ is set to ensure that
 3388 the probability distribution $p_{KN}(w | u)$ sums to one over the vocabulary w .

3389 The idea of modeling versatility by counting contexts may seem heuristic, but there is
 3390 an elegant theoretical justification from Bayesian nonparametrics (Teh, 2006). Kneser-Ney
 3391 smoothing on n -grams was the dominant language modeling technique before the arrival
 3392 of neural language models.

3393 **6.3 Recurrent neural network language models**

3394 N -gram language models have been largely supplanted by neural networks. These mod-
 3395 els do not make the n -gram assumption of restricted context; indeed, they can incorporate
 3396 arbitrarily distant contextual information, while remaining computationally and statisti-
 3397 cally tractable.

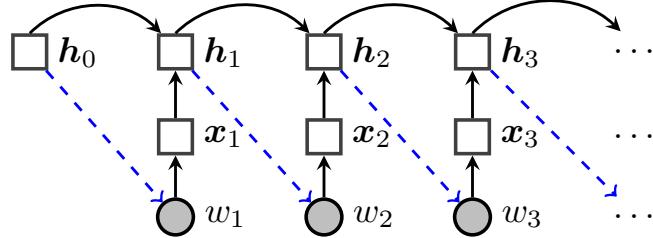


Figure 6.1: The recurrent neural network language model, viewed as an “unrolled” computation graph. Solid lines indicate direct computation, dotted blue lines indicate probabilistic dependencies, circles indicate random variables, and squares indicate computation nodes.

3398 The first insight behind neural language models is to treat word prediction as a *dis-*
 3399 *criminative* learning task.² The goal is to compute the probability $p(w | u)$, where $w \in \mathcal{V}$ is
 3400 a word, and u is the context, which depends on the previous words. Rather than directly
 3401 estimating the word probabilities from (smoothed) relative frequencies, we can treat
 3402 language modeling as a machine learning problem, and estimate parameters that maxi-
 3403 mize the log conditional probability of a corpus.

3404 The second insight is to reparametrize the probability distribution $p(w | u)$ as a func-
 3405 tion of two dense K -dimensional numerical vectors, $\beta_w \in \mathbb{R}^K$, and $v_u \in \mathbb{R}^K$,

$$p(w | u) = \frac{\exp(\beta_w \cdot v_u)}{\sum_{w' \in \mathcal{V}} \exp(\beta_{w'} \cdot v_u)}, \quad [6.25]$$

3406 where $\beta_w \cdot v_u$ represents a dot product. As usual, the denominator ensures that the prob-
 3407 ability distribution is properly normalized. This vector of probabilities is equivalent to
 3408 applying the **softmax** transformation (see § 3.1) to the vector of dot-products,

$$p(\cdot | u) = \text{SoftMax}([\beta_1 \cdot v_u, \beta_2 \cdot v_u, \dots, \beta_V \cdot v_u]). \quad [6.26]$$

The word vectors β_w are parameters of the model, and are estimated directly. The context vectors v_u can be computed in various ways, depending on the model. A simple but effective neural language model can be built from a **recurrent neural network** (RNN; Mikolov et al., 2010). The basic idea is to recurrently update the context vectors while moving through the sequence. Let h_m represent the contextual information at position m

²This idea predates neural language models (e.g., Rosenfeld, 1996; Roark et al., 2007).

in the sequence. RNN language models are defined,

$$\mathbf{x}_m \triangleq \phi_{w_m} \quad [6.27]$$

$$\mathbf{h}_m = \text{RNN}(\mathbf{x}_m, \mathbf{h}_{m-1}) \quad [6.28]$$

$$p(w_{m+1} | w_1, w_2, \dots, w_m) = \frac{\exp(\beta_{w_{m+1}} \cdot \mathbf{h}_m)}{\sum_{w' \in \mathcal{V}} \exp(\beta_{w'} \cdot \mathbf{h}_m)}, \quad [6.29]$$

3409 where ϕ is a matrix of **word embeddings**, and \mathbf{x}_m denotes the embedding for word w_m .
 3410 The conversion of w_m to \mathbf{x}_m is sometimes known as a **lookup layer**, because we simply
 3411 lookup the embeddings for each word in a table; see § 3.2.4.

3412 The **Elman unit** defines a simple recurrent operation (Elman, 1990),

$$\text{RNN}(\mathbf{x}_m, \mathbf{h}_{m-1}) \triangleq g(\Theta \mathbf{h}_{m-1} + \mathbf{x}_m), \quad [6.30]$$

3413 where $\Theta \in \mathbb{R}^{K \times K}$ is the recurrence matrix and g is a non-linear transformation function,
 3414 often defined as the elementwise hyperbolic tangent \tanh (see § 3.1).³ The \tanh acts as a
 3415 **squashing function**, ensuring that each element of \mathbf{h}_m is constrained to the range $[-1, 1]$.

3416 Although each w_m depends on only the context vector \mathbf{h}_{m-1} , this vector is in turn
 3417 influenced by *all* previous tokens, w_1, w_2, \dots, w_{m-1} , through the recurrence operation: w_1
 3418 affects \mathbf{h}_1 , which affects \mathbf{h}_2 , and so on, until the information is propagated all the way to
 3419 \mathbf{h}_{m-1} , and then on to w_m (see Figure 6.1). This is an important distinction from n -gram
 3420 language models, where any information outside the n -word window is ignored. In principle,
 3421 the RNN language model can handle long-range dependencies, such as number
 3422 agreement over long spans of text — although it would be difficult to know where exactly
 3423 in the vector \mathbf{h}_m this information is represented. The main limitation is that information
 3424 is attenuated by repeated application of the squashing function g . **Long short-term**
 3425 **memories** (LSTMs), described below, are a variant of RNNs that address this issue, us-
 3426 ing memory cells to propagate information through the sequence without applying non-
 3427 linearities (Hochreiter and Schmidhuber, 1997).

3428 The denominator in Equation 6.29 is a computational bottleneck, because it involves
 3429 a sum over the entire vocabulary. One solution is to use a **hierarchical softmax** function,
 3430 which computes the sum more efficiently by organizing the vocabulary into a tree (Mikolov
 3431 et al., 2011). Another strategy is to optimize an alternative metric, such as **noise-contrastive**
 3432 **estimation** (Gutmann and Hyvärinen, 2012), which learns by distinguishing observed in-
 3433 stances from artificial instances generated from a noise distribution (Mnih and Teh, 2012).
 3434 Both of these strategies are described in § 14.5.3.

³In the original Elman network, the sigmoid function was used in place of \tanh . For an illuminating mathematical discussion of the advantages and disadvantages of various nonlinearities in recurrent neural networks, see the lecture notes from Cho (2015).

3435 **6.3.1 Backpropagation through time**

3436 The recurrent neural network language model has the following parameters:

- 3437 • $\phi_i \in \mathbb{R}^K$, the “input” word vectors (these are sometimes called **word embeddings**,
3438 since each word is embedded in a K -dimensional space; see chapter 14);
- 3439 • $\beta_i \in \mathbb{R}^K$, the “output” word vectors;
- 3440 • $\Theta \in \mathbb{R}^{K \times K}$, the recurrence operator;
- 3441 • \mathbf{h}_0 , the initial state.

3442 Each of these parameters can be estimated by formulating an objective function over the
3443 training corpus, $L(\mathbf{w})$, and then applying backpropagation to obtain gradients on the
3444 parameters from a minibatch of training examples (see § 3.3.1). Gradient-based updates
3445 can be computed from an online learning algorithm such as stochastic gradient descent
3446 (see § 2.6.2).

3447 The application of backpropagation to recurrent neural networks is known as **back-**
3448 **propagation through time**, because the gradients on units at time m depend in turn on the
3449 gradients of units at earlier times $n < m$. Let ℓ_{m+1} represent the negative log-likelihood
3450 of word $m + 1$,

$$\ell_{m+1} = -\log p(w_{m+1} | w_1, w_2, \dots, w_m). \quad [6.31]$$

We require the gradient of this loss with respect to each parameter, such as $\theta_{k,k'}$, an individual element in the recurrence matrix Θ . Since the loss depends on the parameters only through \mathbf{h}_m , we can apply the chain rule of differentiation,

$$\frac{\partial \ell_{m+1}}{\partial \theta_{k,k'}} = \frac{\partial \ell_{m+1}}{\partial \mathbf{h}_m} \frac{\partial \mathbf{h}_m}{\partial \theta_{k,k'}}. \quad [6.32]$$

The vector \mathbf{h}_m depends on Θ in several ways. First, \mathbf{h}_m is computed by multiplying Θ by the previous state \mathbf{h}_{m-1} . But the previous state \mathbf{h}_{m-1} also depends on Θ :

$$\mathbf{h}_m = g(\mathbf{x}_m, \mathbf{h}_{m-1}) \quad [6.33]$$

$$\frac{\partial h_{m,k}}{\partial \theta_{k,k'}} = g'(\mathbf{x}_{m,k} + \boldsymbol{\theta}_k \cdot \mathbf{h}_{m-1})(h_{m-1,k'} + \boldsymbol{\theta}_k \cdot \frac{\partial \mathbf{h}_{m-1}}{\partial \theta_{k,k'}}), \quad [6.34]$$

3451 where g' is the local derivative of the nonlinear function g . The key point in this equation
3452 is that the derivative $\frac{\partial \mathbf{h}_m}{\partial \theta_{k,k'}}$ depends on $\frac{\partial \mathbf{h}_{m-1}}{\partial \theta_{k,k'}}$, which will depend in turn on $\frac{\partial \mathbf{h}_{m-2}}{\partial \theta_{k,k'}}$, and
3453 so on, until reaching the initial state \mathbf{h}_0 .

3454 Each derivative $\frac{\partial \mathbf{h}_m}{\partial \theta_{k,k'}}$ will be reused many times: it appears in backpropagation from
3455 the loss ℓ_m , but also in all subsequent losses $\ell_{n>m}$. Neural network toolkits such as
3456 Torch (Collobert et al., 2011) and DyNet (Neubig et al., 2017) compute the necessary

3457 derivatives automatically, and cache them for future use. An important distinction from
 3458 the feedforward neural networks considered in chapter 3 is that the size of the computa-
 3459 tion graph is not fixed, but varies with the length of the input. This poses difficulties for
 3460 toolkits that are designed around static computation graphs, such as TensorFlow (Abadi
 3461 et al., 2016).⁴

3462 **6.3.2 Hyperparameters**

3463 The RNN language model has several hyperparameters that must be tuned to ensure good
 3464 performance. The model capacity is controlled by the size of the word and context vectors
 3465 K , which play a role that is somewhat analogous to the size of the n -gram context. For
 3466 datasets that are large with respect to the vocabulary (i.e., there is a large token-to-type
 3467 ratio), we can afford to estimate a model with a large K , which enables more subtle dis-
 3468 tinctions between words and contexts. When the dataset is relatively small, then K must
 3469 be smaller too, or else the model may “memorize” the training data, and fail to generalize.
 3470 Unfortunately, this general advice has not yet been formalized into any concrete formula
 3471 for choosing K , and trial-and-error is still necessary. Overfitting can also be prevented by
 3472 **dropout**, which involves randomly setting some elements of the computation to zero (Sri-
 3473 vastava et al., 2014), forcing the learner not to rely too much on any particular dimension
 3474 of the word or context vectors. The dropout rate must also be tuned on development data.

3475 **6.3.3 Gated recurrent neural networks**

3476 In principle, recurrent neural networks can propagate information across infinitely long
 3477 sequences. But in practice, repeated applications of the nonlinear recurrence function
 3478 causes this information to be quickly attenuated. The same problem affects learning: back-
 3479 propagation can lead to **vanishing gradients** that decay to zero, or **exploding gradients**
 3480 that increase towards infinity (Bengio et al., 1994). The exploding gradient problem can
 3481 be addressed by clipping gradients at some maximum value (Pascanu et al., 2013). The
 3482 other issues must be addressed by altering the model itself.

3483 The **long short-term memory** (LSTM; Hochreiter and Schmidhuber, 1997) is a popular
 3484 variant of RNNs that is more robust to these problems. This model augments the hidden
 3485 state \mathbf{h}_m with a **memory cell** c_m . The value of the memory cell at each time m is a gated
 3486 sum of two quantities: its previous value c_{m-1} , and an “update” \tilde{c}_m , which is computed
 3487 from the current input x_m and the previous hidden state \mathbf{h}_{m-1} . The next state \mathbf{h}_m is then
 3488 computed from the memory cell. Because the memory cell is not passed through a non-
 3489 linear squashing function during the update, it is possible for information to propagate
 3490 through the network over long distances.

⁴See <https://www.tensorflow.org/tutorials/recurrent> (retrieved Feb 8, 2018).

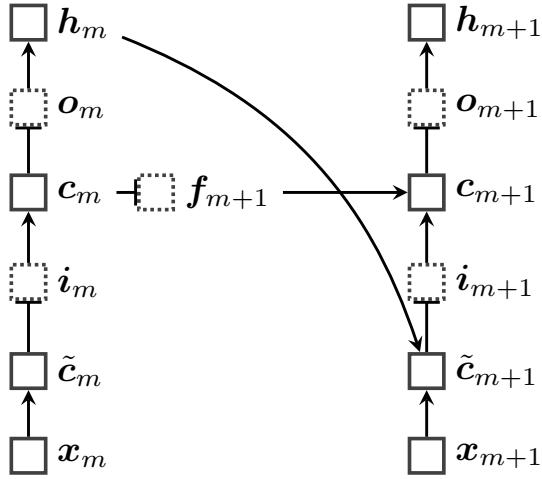


Figure 6.2: The long short-term memory (LSTM) architecture. Gates are shown in boxes with dotted edges. In an LSTM language model, each h_m would be used to predict the next word w_{m+1} .

The gates are functions of the input and previous hidden state. They are computed from elementwise sigmoid activations, $\sigma(x) = (1 + \exp(-x))^{-1}$, ensuring that their values will be in the range $[0, 1]$. They can therefore be viewed as soft, differentiable logic gates. The LSTM architecture is shown in Figure 6.2, and the complete update equations are:

$$f_{m+1} = \sigma(\Theta^{(h \rightarrow f)} h_m + \Theta^{(x \rightarrow f)} x_{m+1} + b_f) \quad \text{forget gate} \quad [6.35]$$

$$i_{m+1} = \sigma(\Theta^{(h \rightarrow i)} h_m + \Theta^{(x \rightarrow i)} x_{m+1} + b_i) \quad \text{input gate} \quad [6.36]$$

$$\tilde{c}_{m+1} = \tanh(\Theta^{(h \rightarrow c)} h_m + \Theta^{(x \rightarrow c)} x_{m+1}) \quad \text{update candidate} \quad [6.37]$$

$$c_{m+1} = f_{m+1} \odot c_m + i_{m+1} \odot \tilde{c}_{m+1} \quad \text{memory cell update} \quad [6.38]$$

$$o_{m+1} = \sigma(\Theta^{(h \rightarrow o)} h_m + \Theta^{(x \rightarrow o)} x_{m+1} + b_o) \quad \text{output gate} \quad [6.39]$$

$$h_{m+1} = o_{m+1} \odot \tanh(c_{m+1}) \quad \text{output.} \quad [6.40]$$

3491 The operator \odot is an elementwise (Hadamard) product. Each gate is controlled by a vec-
 3492 tor of weights, which parametrize the previous hidden state (e.g., $\Theta^{(h \rightarrow f)}$) and the current
 3493 input (e.g., $\Theta^{(x \rightarrow f)}$), plus a vector offset (e.g., b_f). The overall operation can be infor-
 3494 mally summarized as $(h_m, c_m) = \text{LSTM}(x_m, (h_{m-1}, c_{m-1}))$, with (h_m, c_m) representing
 3495 the LSTM state after reading token m .

3496 The LSTM outperforms standard recurrent neural networks across a wide range of
 3497 problems. It was first used for language modeling by Sundermeyer et al. (2012), but can
 3498 be applied more generally: the vector h_m can be treated as a complete representation of

3499 the input sequence up to position m , and can be used for any labeling task on a sequence
 3500 of tokens, as we will see in the next chapter.

3501 There are several LSTM variants, of which the Gated Recurrent Unit (Cho et al., 2014)
 3502 is one of the more well known. Many software packages implement a variety of RNN
 3503 architectures, so choosing between them is simple from a user’s perspective. Jozefowicz
 3504 et al. (2015) provide an empirical comparison of various modeling choices circa 2015.

3505 6.4 Evaluating language models

3506 Language modeling is not usually an application in itself: language models are typically
 3507 components of larger systems, and they would ideally be evaluated **extrinsically**. This
 3508 means evaluating whether the language model improves performance on the application
 3509 task, such as machine translation or speech recognition. But this is often hard to do, and
 3510 depends on details of the overall system which may be irrelevant to language modeling.
 3511 In contrast, **intrinsic evaluation** is task-neutral. Better performance on intrinsic metrics
 3512 may be expected to improve extrinsic metrics across a variety of tasks, but there is always
 3513 the risk of over-optimizing the intrinsic metric. This section discusses some intrinsic met-
 3514 rics, but keep in mind the importance of performing extrinsic evaluations to ensure that
 3515 intrinsic performance gains carry over to real applications.

3516 6.4.1 Held-out likelihood

The goal of probabilistic language models is to accurately measure the probability of sequences of word tokens. Therefore, an intrinsic evaluation metric is the likelihood that the language model assigns to **held-out data**, which is not used during training. Specifically, we compute,

$$\ell(\mathbf{w}) = \sum_{m=1}^M \log p(w_m | w_{m-1}, \dots, w_1), \quad [6.41]$$

3517 treating the entire held-out corpus as a single stream of tokens.

3518 Typically, unknown words are mapped to the $\langle \text{UNK} \rangle$ token. This means that we have
 3519 to estimate some probability for $\langle \text{UNK} \rangle$ on the training data. One way to do this is to fix
 3520 the vocabulary \mathcal{V} to the $V - 1$ words with the highest counts in the training data, and then
 3521 convert all other tokens to $\langle \text{UNK} \rangle$. Other strategies for dealing with out-of-vocabulary
 3522 terms are discussed in § 6.5.

3523 **6.4.2 Perplexity**

Held-out likelihood is usually presented as **perplexity**, which is a deterministic transformation of the log-likelihood into an information-theoretic quantity,

$$\text{Perplex}(\mathbf{w}) = 2^{-\frac{\ell(\mathbf{w})}{M}}, \quad [6.42]$$

3524 where M is the total number of tokens in the held-out corpus.

3525 Lower perplexities correspond to higher likelihoods, so lower scores are better on this
3526 metric — it is better to be less perplexed. Here are some special cases:

- 3527 • In the limit of a perfect language model, probability 1 is assigned to the held-out
3528 corpus, with $\text{Perplex}(\mathbf{w}) = 2^{-\frac{1}{M} \log_2 1} = 2^0 = 1$.
- 3529 • In the opposite limit, probability zero is assigned to the held-out corpus, which cor-
3530 responds to an infinite perplexity, $\text{Perplex}(\mathbf{w}) = 2^{-\frac{1}{M} \log_2 0} = 2^\infty = \infty$.
- Assume a uniform, unigram model in which $p(w_i) = \frac{1}{V}$ for all words in the vocab-
3531 ular. Then,

$$\begin{aligned} \log_2(\mathbf{w}) &= \sum_{m=1}^M \log_2 \frac{1}{V} = - \sum_{m=1}^M \log_2 V = -M \log_2 V \\ \text{Perplex}(\mathbf{w}) &= 2^{\frac{1}{M} M \log_2 V} \\ &= 2^{\log_2 V} \\ &= V. \end{aligned}$$

3531 This is the “worst reasonable case” scenario, since you could build such a language
3532 model without even looking at the data.

3533 In practice, language models tend to give perplexities in the range between 1 and V .
3534 A small benchmark dataset is the **Penn Treebank**, which contains roughly a million to-
3535 kens; its vocabulary is limited to 10,000 words, with all other tokens mapped a special
3536 $\langle \text{UNK} \rangle$ symbol. On this dataset, a well-smoothed 5-gram model achieves a perplexity of
3537 141 (Mikolov and Zweig, Mikolov and Zweig), and an LSTM language model achieves
3538 perplexity of roughly 80 (Zaremba, Sutskever, and Vinyals, Zaremba et al.). Various en-
3539 hancements to the LSTM architecture can bring the perplexity below 60 (Merity et al.,
3540 2018). A larger-scale language modeling dataset is the 1B Word Benchmark (Chelba et al.,
3541 2013), which contains text from Wikipedia. On this dataset, a perplexities of around 25
3542 can be obtained by averaging together multiple LSTM language models (Jozefowicz et al.,
3543 2016).

3544 **6.5 Out-of-vocabulary words**

3545 So far, we have assumed a **closed-vocabulary** setting — the vocabulary \mathcal{V} is assumed to be
 3546 a finite set. In realistic application scenarios, this assumption may not hold. Consider, for
 3547 example, the problem of translating newspaper articles. The following sentence appeared
 3548 in a Reuters article on January 6, 2017:⁵

3549 The report said U.S. intelligence agencies believe Russian military intelligence,
 3550 the **GRU**, used intermediaries such as **WikiLeaks**, **DCLeaks.com** and the **Guc-**
 3551 **cifer** 2.0 "persona" to release emails...

3552 Suppose that you trained a language model on the Gigaword corpus,⁶ which was released
 3553 in 2003. The bolded terms either did not exist at this date, or were not widely known; they
 3554 are unlikely to be in the vocabulary. The same problem can occur for a variety of other
 3555 terms: new technologies, previously unknown individuals, new words (e.g., *hashtag*), and
 3556 numbers.

3557 One solution is to simply mark all such terms with a special token, $\langle \text{UNK} \rangle$. While
 3558 training the language model, we decide in advance on the vocabulary (often the K most
 3559 common terms), and mark all other terms in the training data as $\langle \text{UNK} \rangle$. If we do not want
 3560 to determine the vocabulary size in advance, an alternative approach is to simply mark
 3561 the first occurrence of each word type as $\langle \text{UNK} \rangle$.

3562 But it often better to make distinctions about the likelihood of various unknown words.
 3563 This is particularly important in languages that have rich morphological systems, with
 3564 many inflections for each word. For example, Portuguese is only moderately complex
 3565 from a morphological perspective, yet each verb has dozens of inflected forms (see Fig-
 3566 ure 4.3b). In such languages, there will be many word types that we do not encounter in a
 3567 corpus, which are nonetheless predictable from the morphological rules of the language.
 3568 To use a somewhat contrived English example, if *transfenestrate* is in the vocabulary, our
 3569 language model should assign a non-zero probability to the past tense *transfenestrated*,
 3570 even if it does not appear in the training data.

3571 One way to accomplish this is to supplement word-level language models with **character-**
 3572 **level language models**. Such models can use n -grams or RNNs, but with a fixed vocab-
 3573 uary equal to the set of ASCII or Unicode characters. For example, Ling et al. (2015)
 3574 propose an LSTM model over characters, and Kim (2014) employ a convolutional neural
 3575 network. A more linguistically motivated approach is to segment words into meaningful
 3576 subword units, known as **morphemes** (see chapter 9). For example, Botha and Blunsom

⁵Bayoumy, Y. and Strobel, W. (2017, January 6). U.S. intel report: Putin directed cy-
 ber campaign to help Trump. *Reuters*. Retrieved from <http://www.reuters.com/article/us-usa-russia-cyber-idUSKBN14Q1T8> on January 7, 2017.

⁶<https://catalog.ldc.upenn.edu/LDC2003T05>

3577 (2014) induce vector representations for morphemes, which they build into a log-bilinear
 3578 language model; Bhatia et al. (2016) incorporate morpheme vectors into an LSTM.

3579 Additional resources

3580 A variety of neural network architectures have been applied to language modeling. No-
 3581 table earlier non-recurrent architectures include the neural probabilistic language model (Ben-
 3582 gio et al., 2003) and the log-bilinear language model (Mnih and Hinton, 2007). Much more
 3583 detail on these models can be found in the text by Goodfellow et al. (2016).

3584 Exercises

- 3585 1. Prove that n -gram language models give valid probabilities if the n -gram probabili-
 3586 ties are valid. Specifically, assume that,

$$\sum_{w_m}^{\mathcal{V}} p(w_m | w_{m-1}, w_{m-2}, \dots, w_{m-n+1}) = 1 \quad [6.43]$$

3587 for all contexts $(w_{m-1}, w_{m-2}, \dots, w_{m-n+1})$. Prove that $\sum_w p_n(w) = 1$ for all $w \in \mathcal{V}^*$,
 3588 where p_n is the probability under an n -gram language model. Your proof should
 3589 proceed by induction. You should handle the start-of-string case $p(w_1 | \underbrace{\square, \dots, \square}_{n-1})$,

3590 but you need not handle the end-of-string token.

- 3591 2. First, show that RNN language models are valid using a similar proof technique to
 3592 the one in the previous problem.

3593 Next, let $p_r(w)$ indicate the probability of w under RNN r . An ensemble of RNN
 3594 language models computes the probability,

$$p(w) = \frac{1}{R} \sum_{r=1}^R p_r(w). \quad [6.44]$$

3595 Does an ensemble of RNN language models compute a valid probability?

- 3596 3. Consider a unigram language model over a vocabulary of size V . Suppose that a
 3597 word appears m times in a corpus with M tokens in total. With Lidstone smoothing
 3598 of α , for what values of m is the smoothed probability greater than the unsmoothed
 3599 probability?
- 3600 4. Consider a simple language in which each token is drawn from the vocabulary \mathcal{V}
 3601 with probability $\frac{1}{V}$, independent of all other tokens.

- Given a corpus of size M , what is the expectation of the fraction of all possible bigrams that have zero count? You may assume V is large enough that $\frac{1}{V} \approx \frac{1}{V-1}$.
5. Continuing the previous problem, determine the value of M such that the fraction of bigrams with zero count is at most $\epsilon \in (0, 1)$. As a hint, you may use the approximation $\ln(1 + \alpha) \approx \alpha$ for $\alpha \approx 0$.
 6. In real languages, word probabilities are neither uniform nor independent. Assume that word probabilities are independent but not uniform, so that in general $p(w) \neq \frac{1}{V}$. Prove that the expected fraction of unseen bigrams will be higher than in the IID case.
 7. Consider a recurrent neural network with a single hidden unit and a sigmoid activation, $h_m = \sigma(\theta h_{m-1} + x_m)$. Prove that if $|\theta| < 1$, then the gradient $\frac{\partial h_m}{\partial h_{m-k}}$ goes to zero as $k \rightarrow \infty$.⁷
 8. **Zipf's law** states that if the word types in a corpus are sorted by frequency, then the frequency of the word at rank r is proportional to r^{-s} , where s is a free parameter, usually around 1. (Another way to view Zipf's law is that a plot of log frequency against log rank will be linear.) Solve for s using the counts of the first and second most frequent words, c_1 and c_2 .
 9. Download the wikitext-2 dataset.⁸ Read in the training data and compute word counts. Estimate the Zipf's law coefficient by,

$$\hat{s} = \exp \left(\frac{(\log r) \cdot (\log c)}{\|\log r\|_2^2} \right), \quad [6.45]$$

where $r = [1, 2, 3, \dots]$ is the vector of ranks of all words in the corpus, and $c = [c_1, c_2, c_3, \dots]$ is the vector of counts of all words in the corpus, sorted in descending order.

Make a log-log plot of the observed counts, and the expected counts according to Zipf's law. The sum $\sum_{r=1}^{\infty} r^s = \zeta(s)$ is the Riemann zeta function, available in python's `scipy` library as `scipy.special.zeta`.

10. Using the Pytorch library, train an LSTM language model from the Wikitext training corpus. After each epoch of training, compute its perplexity on the Wikitext validation corpus. Stop training when the perplexity stops improving.

⁷This proof generalizes to vector hidden units by considering the largest eigenvector of the matrix Θ (Pascanu et al., 2013).

⁸Available at https://github.com/pytorch/examples/tree/master/word_language_model/data/wikitext-2 in September 2018. The dataset is already tokenized, and already replaces rare words with `<UNK>`, so no preprocessing is necessary.

3630 Chapter 7

3631 Sequence labeling

3632 The goal of sequence labeling is to assign tags to words, or more generally, to assign
3633 discrete labels to discrete elements in a sequence. There are many applications of se-
3634 quence labeling in natural language processing, and chapter 8 presents an overview. For
3635 now, we'll focus on the classic problem of **part-of-speech tagging**, which requires tagging
3636 each word by its grammatical category. Coarse-grained grammatical categories include
3637 **NOUNs**, which describe things, properties, or ideas, and **VERBS**, which describe actions
3638 and events. Consider a simple input:

3639 (7.1) They can fish.

3640 A dictionary of coarse-grained part-of-speech tags might include **NOUN** as the only valid
3641 tag for *they*, but both **NOUN** and **VERB** as potential tags for *can* and *fish*. An accurate se-
3642 quence labeling algorithm should select the verb tag for both *can* and *fish* in (7.1), but it
3643 should select noun for the same two words in the phrase *can of fish*.

3644 7.1 Sequence labeling as classification

One way to solve a tagging problem is to turn it into a classification problem. Let $f((\mathbf{w}, m), y)$ indicate the feature function for tag y at position m in the sequence $\mathbf{w} = (w_1, w_2, \dots, w_M)$. A simple tagging model would have a single base feature, the word itself:

$$f((\mathbf{w} = \text{they can fish}, m = 1), \text{N}) = (\text{they}, \text{N}) \quad [7.1]$$

$$f((\mathbf{w} = \text{they can fish}, m = 2), \text{V}) = (\text{can}, \text{V}) \quad [7.2]$$

$$f((\mathbf{w} = \text{they can fish}, m = 3), \text{V}) = (\text{fish}, \text{V}). \quad [7.3]$$

3645 Here the feature function takes three arguments as input: the sentence to be tagged (e.g.,
3646 *they can fish*), the proposed tag (e.g., N or V), and the index of the token to which this tag

3647 is applied. This simple feature function then returns a single feature: a tuple including
 3648 the word to be tagged and the tag that has been proposed. If the vocabulary size is V
 3649 and the number of tags is K , then there are $V \times K$ features. Each of these features must
 3650 be assigned a weight. These weights can be learned from a labeled dataset using a clas-
 3651 sification algorithm such as perceptron, but this isn't necessary in this case: it would be
 3652 equivalent to define the classification weights directly, with $\theta_{w,y} = 1$ for the tag y most
 3653 frequently associated with word w , and $\theta_{w,y} = 0$ for all other tags.

However, it is easy to see that this simple classification approach cannot correctly tag both *they can fish* and *can of fish*, because *can* and *fish* are grammatically ambiguous. To handle both of these cases, the tagger must rely on context, such as the surrounding words. We can build context into the feature set by incorporating the surrounding words as additional features:

$$\begin{aligned} f((\mathbf{w} = \text{they can fish}, 1), \mathbf{N}) = & \{(w_m = \text{they}, y_m = \mathbf{N}), \\ & (w_{m-1} = \square, y_m = \mathbf{N}), \\ & (w_{m+1} = \text{can}, y_m = \mathbf{N})\} \end{aligned} \quad [7.4]$$

$$\begin{aligned} f((\mathbf{w} = \text{they can fish}, 2), \mathbf{V}) = & \{(w_m = \text{can}, y_m = \mathbf{V}), \\ & (w_{m-1} = \text{they}, y_m = \mathbf{V}), \\ & (w_{m+1} = \text{fish}, y_m = \mathbf{V})\} \end{aligned} \quad [7.5]$$

$$\begin{aligned} f((\mathbf{w} = \text{they can fish}, 3), \mathbf{V}) = & \{(w_m = \text{fish}, y_m = \mathbf{V}), \\ & (w_{m-1} = \text{can}, y_m = \mathbf{V}), \\ & (w_{m+1} = \blacksquare, y_m = \mathbf{V})\}. \end{aligned} \quad [7.6]$$

3654 These features contain enough information that a tagger should be able to choose the
 3655 right tag for the word *fish*: words that come after *can* are likely to be verbs, so the feature
 3656 $(w_{m-1} = \text{can}, y_m = \mathbf{V})$ should have a large positive weight.

3657 However, even with this enhanced feature set, it may be difficult to tag some se-
 3658 quences correctly. One reason is that there are often relationships between the tags them-
 3659 selves. For example, in English it is relatively rare for a verb to follow another verb —
 3660 particularly if we differentiate MODAL verbs like *can* and *should* from more typical verbs,
 3661 like *give*, *transcend*, and *befuddle*. We would like to incorporate preferences against tag se-
 3662 quences like VERB-VERB, and in favor of tag sequences like NOUN-VERB. The need for
 3663 such preferences is best illustrated by a **garden path sentence**:

3664 (7.2) The old man the boat.

3665 Grammatically, the word *the* is a DETERMINER. When you read the sentence, what
 3666 part of speech did you first assign to *old*? Typically, this word is an ADJECTIVE — abbrevi-
 3667 ated as J — which is a class of words that modify nouns. Similarly, *man* is usually a noun.
 3668 The resulting sequence of tags is D J N D N. But this is a mistaken “garden path” inter-
 3669 pretation, which ends up leading nowhere. It is unlikely that a determiner would directly

follow a noun,¹ and it is particularly unlikely that the entire sentence would lack a verb. The only possible verb in (7.2) is the word *man*, which can refer to the act of maintaining and piloting something — often boats. But if *man* is tagged as a verb, then *old* is seated between a determiner and a verb, and must be a noun. And indeed, adjectives often have a second interpretation as nouns when used in this way (e.g., *the young*, *the restless*). This reasoning, in which the labeling decisions are intertwined, cannot be applied in a setting where each tag is produced by an independent classification decision.

7.2 Sequence labeling as structure prediction

As an alternative, think of the entire sequence of tags as a label itself. For a given sequence of words $\mathbf{w} = (w_1, w_2, \dots, w_M)$, there is a set of possible taggings $\mathcal{Y}(\mathbf{w}) = \mathcal{Y}^M$, where $\mathcal{Y} = \{\text{N, V, D, ...}\}$ refers to the set of individual tags, and \mathcal{Y}^M refers to the set of tag sequences of length M . We can then treat the sequence labeling problem as a classification problem in the label space $\mathcal{Y}(\mathbf{w})$,

$$\hat{\mathbf{y}} = \underset{\mathbf{y} \in \mathcal{Y}(\mathbf{w})}{\operatorname{argmax}} \Psi(\mathbf{w}, \mathbf{y}), \quad [7.7]$$

where $\mathbf{y} = (y_1, y_2, \dots, y_M)$ is a sequence of M tags, and Ψ is a scoring function on pairs of sequences, $V^M \times \mathcal{Y}^M \rightarrow \mathbb{R}$. Such a function can include features that capture the relationships between tagging decisions, such as the preference that determiners not follow nouns, or that all sentences have verbs.

Given that the label space is exponentially large in the length of the sequence M , can it ever be practical to perform tagging in this way? The problem of making a series of interconnected labeling decisions is known as **inference**. Because natural language is full of interrelated grammatical structures, inference is a crucial aspect of natural language processing. In English, it is not unusual to have sentences of length $M = 20$; part-of-speech tag sets vary in size from 10 to several hundred. Taking the low end of this range, we have $|\mathcal{Y}(\mathbf{w}_{1:M})| \approx 10^{20}$, one hundred billion billion possible tag sequences. Enumerating and scoring each of these sequences would require an amount of work that is exponential in the sequence length, so inference is intractable.

However, the situation changes when we restrict the scoring function. Suppose we choose a function that decomposes into a sum of local parts,

$$\Psi(\mathbf{w}, \mathbf{y}) = \sum_{m=1}^{M+1} \psi(\mathbf{w}, y_m, y_{m-1}, m), \quad [7.8]$$

where each $\psi(\cdot)$ scores a local part of the tag sequence. Note that the sum goes up to $M+1$, so that we can include a score for a special end-of-sequence tag, $\psi(\mathbf{w}_{1:M}, \diamond, y_M, M+1)$. We also define a special tag to begin the sequence, $y_0 \triangleq \diamond$.

¹The main exception occurs with ditransitive verbs, such as *They gave the winner a trophy*.

3699 In a linear model, local scoring function can be defined as a dot product of weights
 3700 and features,

$$\psi(\mathbf{w}_{1:M}, y_m, y_{m-1}, m) = \boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{w}, y_m, y_{m-1}, m). \quad [7.9]$$

3701 The feature vector \mathbf{f} can consider the entire input \mathbf{w} , and can look at pairs of adjacent
 3702 tags. This is a step up from per-token classification: the weights can assign low scores
 3703 to infelicitous tag pairs, such as noun-determiner, and high scores for frequent tag pairs,
 3704 such as determiner-noun and noun-verb.

In the example *they can fish*, a minimal feature function would include features for word-tag pairs (sometimes called **emission features**) and tag-tag pairs (sometimes called **transition features**):

$$\begin{aligned} \mathbf{f}(\mathbf{w} = \text{they can fish}, \mathbf{y} = \text{N V V}) &= \sum_{m=1}^{M+1} \mathbf{f}(\mathbf{w}, y_m, y_{m-1}, m) \\ &= \mathbf{f}(\mathbf{w}, \text{N}, \diamond, 1) \\ &\quad + \mathbf{f}(\mathbf{w}, \text{V}, \text{N}, 2) \\ &\quad + \mathbf{f}(\mathbf{w}, \text{V}, \text{V}, 3) \\ &\quad + \mathbf{f}(\mathbf{w}, \blacklozenge, \text{V}, 4) \end{aligned} \quad [7.10]$$

$$\begin{aligned} &= (w_m = \text{they}, y_m = \text{N}) + (y_m = \text{N}, y_{m-1} = \diamond) \\ &\quad + (w_m = \text{can}, y_m = \text{V}) + (y_m = \text{V}, y_{m-1} = \text{N}) \\ &\quad + (w_m = \text{fish}, y_m = \text{V}) + (y_m = \text{V}, y_{m-1} = \text{V}) \\ &\quad + (y_m = \blacklozenge, y_{m-1} = \text{V}). \end{aligned} \quad [7.11]$$

3705 There are seven active features for this example: one for each word-tag pair, and one
 3706 for each tag-tag pair, including a final tag $y_{M+1} = \blacklozenge$. These features capture the two main
 3707 sources of information for part-of-speech tagging in English: which tags are appropriate
 3708 for each word, and which tags tend to follow each other in sequence. Given appropriate
 3709 weights for these features, taggers can achieve high accuracy, even for difficult cases like
 3710 *the old man the boat*. We will now discuss how this restricted scoring function enables
 3711 efficient inference, through the **Viterbi algorithm** (Viterbi, 1967).

³⁷¹² **7.3 The Viterbi algorithm**

By decomposing the scoring function into a sum of local parts, it is possible to rewrite the tagging problem as follows:

$$\hat{\mathbf{y}} = \operatorname{argmax}_{\mathbf{y} \in \mathcal{Y}(\mathbf{w})} \Psi(\mathbf{w}, \mathbf{y}) \quad [7.13]$$

$$= \operatorname{argmax}_{\mathbf{y}_{1:M}} \sum_{m=1}^{M+1} \psi(\mathbf{w}, y_m, y_{m-1}, m) \quad [7.14]$$

$$= \operatorname{argmax}_{\mathbf{y}_{1:M}} \sum_{m=1}^{M+1} s_m(y_m, y_{m-1}), \quad [7.15]$$

³⁷¹³ where the final line simplifies the notation with the shorthand,

$$s_m(y_m, y_{m-1}) \triangleq \psi(\mathbf{w}_{1:M}, y_m, y_{m-1}, m). \quad [7.16]$$

This inference problem can be solved efficiently using **dynamic programming**, an algorithmic technique for reusing work in recurrent computations. We begin by solving an auxiliary problem: rather than finding the best tag sequence, we compute the *score* of the best tag sequence,

$$\max_{\mathbf{y}_{1:M}} \Psi(\mathbf{w}, \mathbf{y}_{1:M}) = \max_{\mathbf{y}_{1:M}} \sum_{m=1}^{M+1} s_m(y_m, y_{m-1}). \quad [7.17]$$

This score involves a maximization over all tag sequences of length M , written $\max_{\mathbf{y}_{1:M}}$. This maximization can be broken into two pieces,

$$\max_{\mathbf{y}_{1:M}} \Psi(\mathbf{w}, \mathbf{y}_{1:M}) = \max_{y_M} \max_{\mathbf{y}_{1:M-1}} \sum_{m=1}^{M+1} s_m(y_m, y_{m-1}). \quad [7.18]$$

Within the sum, only the final term $s_{M+1}(\blacklozenge, y_M)$ depends on y_M , so we can pull this term out of the second maximization,

$$\max_{\mathbf{y}_{1:M}} \Psi(\mathbf{w}, \mathbf{y}_{1:M}) = \left(\max_{y_M} s_{M+1}(\blacklozenge, y_M) \right) + \left(\max_{\mathbf{y}_{1:M-1}} \sum_{m=1}^M s_m(y_m, y_{m-1}) \right). \quad [7.19]$$

The second term in Equation 7.19 has the same form as our original problem, with M replaced by $M-1$. This indicates that the problem can be reformulated as a recurrence. We do this by defining an auxiliary variable called the **Viterbi variable** $v_m(k)$, representing

Algorithm 11 The Viterbi algorithm. Each $s_m(k, k')$ is a local score for tag $y_m = k$ and $y_{m-1} = k'$.

```

for  $k \in \{0, \dots, K\}$  do
     $v_1(k) = s_1(k, \diamond)$ 
for  $m \in \{2, \dots, M\}$  do
    for  $k \in \{0, \dots, K\}$  do
         $v_m(k) = \max_{k'} s_m(k, k') + v_{m-1}(k')$ 
         $b_m(k) = \operatorname{argmax}_{k'} s_m(k, k') + v_{m-1}(k')$ 
     $y_M = \operatorname{argmax}_k s_{M+1}(\diamond, k) + v_M(k)$ 
    for  $m \in \{M-1, \dots, 1\}$  do
         $y_m = b_m(y_{m+1})$ 
return  $\mathbf{y}_{1:M}$ 
```

the score of the best sequence terminating in the tag k :

$$v_m(y_m) \triangleq \max_{\mathbf{y}_{1:m-1}} \sum_{n=1}^m s_n(y_n, y_{n-1}) \quad [7.20]$$

$$= \max_{y_{m-1}} s_m(y_m, y_{m-1}) + \max_{\mathbf{y}_{1:m-2}} \sum_{n=1}^{m-1} s_n(y_n, y_{n-1}) \quad [7.21]$$

$$= \max_{y_{m-1}} s_m(y_m, y_{m-1}) + v_{m-1}(y_{m-1}). \quad [7.22]$$

Each set of Viterbi variables is computed from the local score $s_m(y_m, y_{m-1})$, and from the previous set of Viterbi variables. The initial condition of the recurrence is simply the score for the first tag,

$$v_1(y_1) \triangleq s_1(y_1, \diamond). \quad [7.23]$$

The maximum overall score for the sequence is then the final Viterbi variable,

$$\max_{\mathbf{y}_{1:M}} \Psi(\mathbf{w}_{1:M}, \mathbf{y}_{1:M}) = v_{M+1}(\diamond). \quad [7.24]$$

3714 Thus, the score of the best labeling for the sequence can be computed in a single forward
 3715 sweep: first compute all variables $v_1(\cdot)$ from Equation 7.23, and then compute all variables
 3716 $v_2(\cdot)$ from the recurrence in Equation 7.22, continuing until the final variable $v_{M+1}(\diamond)$.

3717 The Viterbi variables can be arranged in a structure known as a **trellis**, shown in Fig-
 3718 ure 7.1. Each column indexes a token m in the sequence, and each row indexes a tag in
 3719 \mathcal{Y} ; every $v_{m-1}(k)$ is connected to every $v_m(k')$, indicating that $v_m(k')$ is computed from
 3720 $v_{m-1}(k)$. Special nodes are set aside for the start and end states.

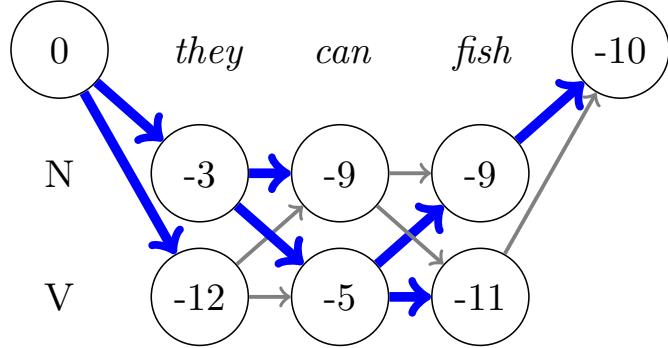


Figure 7.1: The trellis representation of the Viterbi variables, for the example *they can fish*, using the weights shown in Table 7.1.

3721 The original goal was to find the best scoring sequence, not simply to compute its
 3722 score. But by solving the auxiliary problem, we are almost there. Recall that each $v_m(k)$
 3723 represents the score of the best tag sequence ending in that tag k in position m . To compute
 3724 this, we maximize over possible values of y_{m-1} . By keeping track of the “argmax” tag that
 3725 maximizes this choice at each step, we can walk backwards from the final tag, and recover
 3726 the optimal tag sequence. This is indicated in Figure 7.1 by the thick lines, which we trace
 3727 back from the final position. These backward pointers are written $b_m(k)$, indicating the
 3728 optimal tag y_{m-1} on the path to $Y_m = k$.

3729 The complete Viterbi algorithm is shown in Algorithm 11. When computing the initial
 3730 Viterbi variables $v_1(\cdot)$, the special tag \diamond indicates the start of the sequence. When comput-
 3731 ing the final tag Y_M , another special tag, \blacklozenge indicates the end of the sequence. These special
 3732 tags enable the use of transition features for the tags that begin and end the sequence: for
 3733 example, conjunctions are unlikely to end sentences in English, so we would like a low
 3734 score for $s_{M+1}(\blacklozenge, CC)$; nouns are relatively likely to appear at the beginning of sentences,
 3735 so we would like a high score for $s_1(N, \diamond)$, assuming the noun tag is compatible with the
 3736 first word token w_1 .

3737 **Complexity** If there are K tags and M positions in the sequence, then there are $M \times K$
 3738 Viterbi variables to compute. Computing each variable requires finding a maximum over
 3739 K possible predecessor tags. The total time complexity of populating the trellis is there-
 3740 fore $\mathcal{O}(MK^2)$, with an additional factor for the number of active features at each position.
 3741 After completing the trellis, we simply trace the backwards pointers to the beginning of
 3742 the sequence, which takes $\mathcal{O}(M)$ operations.

	<i>they</i>	<i>can</i>	<i>fish</i>	
N	-2	-3	-3	
V	-10	-1	-3	

(a) Weights for emission features.

	N	V	♦
◊	-1	-2	$-\infty$
N	-3	-1	-1
V	-1	-3	-1

(b) Weights for transition features. The “from” tags are on the columns, and the “to” tags are on the rows.

Table 7.1: Feature weights for the example trellis shown in Figure 7.1. Emission weights from \diamond and ♦ are implicitly set to $-\infty$.3743 **7.3.1 Example**

3744 Consider the minimal tagset $\{N, V\}$, corresponding to nouns and verbs. Even in this
 3745 tagset, there is considerable ambiguity: for example, the words *can* and *fish* can each take
 3746 both tags. Of the $2 \times 2 \times 2 = 8$ possible taggings for the sentence *they can fish*, four are
 3747 possible given these possible tags, and two are grammatical.²

3748 The values in the trellis in Figure 7.1 are computed from the feature weights defined in
 3749 Table 7.1. We begin with $v_1(N)$, which has only one possible predecessor, the start tag \diamond .
 3750 This score is therefore equal to $s_1(N, \diamond) = -2 - 1 = -3$, which is the sum of the scores for
 3751 the emission and transition features respectively; the backpointer is $b_1(N) = \diamond$. The score
 3752 for $v_1(V)$ is computed in the same way: $s_1(V, \diamond) = -10 - 2 = -12$, and again $b_1(V) = \diamond$.
 3753 The backpointers are represented in the figure by thick lines.

Things get more interesting at $m = 2$. The score $v_2(N)$ is computed by maximizing over the two possible predecessors,

$$v_2(N) = \max(v_1(N) + s_2(N, N), v_1(V) + s_2(N, V)) \quad [7.25]$$

$$= \max(-3 - 3 - 3, -12 - 3 - 1) = -9 \quad [7.26]$$

$$b_2(N) = N. \quad [7.27]$$

This continues until reaching $v_4(\diamond)$, which is computed as,

$$v_4(\diamond) = \max(v_3(N) + s_4(\diamond, N), v_3(V) + s_4(\diamond, V)) \quad [7.28]$$

$$= \max(-9 + 0 - 1, -11 + 0 - 1) \quad [7.29]$$

$$= -10, \quad [7.30]$$

3754 so $b_4(\diamond) = N$. As there is no emission w_4 , the emission features have scores of zero.

²The tagging *they/N can/V fish/N* corresponds to the scenario of putting fish into cans, or perhaps of firing them.

3755 To compute the optimal tag sequence, we walk backwards from here, next checking
 3756 $b_3(N) = V$, and then $b_2(V) = N$, and finally $b_1(N) = \diamond$. This yields $y = (N, V, N)$, which
 3757 corresponds to the linguistic interpretation of the fishes being put into cans.

3758 **7.3.2 Higher-order features**

3759 The Viterbi algorithm was made possible by a restriction of the scoring function to local
 3760 parts that consider only pairs of adjacent tags. We can think of this as a bigram language
 3761 model over tags. A natural question is how to generalize Viterbi to tag trigrams, which
 3762 would involve the following decomposition:

$$\Psi(\mathbf{w}, \mathbf{y}) = \sum_{m=1}^{M+2} f(\mathbf{w}, y_m, y_{m-1}, y_{m-2}, m), \quad [7.31]$$

3763 where $y_{-1} = \diamond$ and $y_{M+2} = \blacklozenge$.

3764 One solution is to create a new tagset $\mathcal{Y}^{(2)}$ from the Cartesian product of the original
 3765 tagset with itself, $\mathcal{Y}^{(2)} = \mathcal{Y} \times \mathcal{Y}$. The tags in this product space are ordered pairs, rep-
 3766 resenting adjacent tags at the token level: for example, the tag (N, V) would represent a
 3767 noun followed by a verb. Transitions between such tags must be consistent: we can have a
 3768 transition from (N, V) to (V, N) (corresponding to the tag sequence $N V N$), but not from
 3769 (N, V) to (N, N) , which would not correspond to any coherent tag sequence. This con-
 3770 straint can be enforced in feature weights, with $\theta_{((a,b),(c,d))} = -\infty$ if $b \neq c$. The remaining
 3771 feature weights can encode preferences for and against various tag trigrams.

3772 In the Cartesian product tag space, there are K^2 tags, suggesting that the time com-
 3773 plexity will increase to $\mathcal{O}(MK^4)$. However, it is unnecessary to max over predecessor tag
 3774 bigrams that are incompatible with the current tag bigram. By exploiting this constraint,
 3775 it is possible to limit the time complexity to $\mathcal{O}(MK^3)$. The space complexity grows to
 3776 $\mathcal{O}(MK^2)$, since the trellis must store all possible predecessors of each tag. In general, the
 3777 time and space complexity of higher-order Viterbi grows exponentially with the order of
 3778 the tag n -grams that are considered in the feature decomposition.

3779 **7.4 Hidden Markov Models**

3780 The Viterbi sequence labeling algorithm is built on the scores $s_m(y, y')$. We will now
 3781 discuss how these scores can be estimated probabilistically. Recall from § 2.2 that the
 3782 probabilistic Naïve Bayes classifier selects the label y to maximize $p(y | \mathbf{x}) \propto p(y, \mathbf{x})$. In
 3783 probabilistic sequence labeling, our goal is similar: select the tag sequence that maximizes
 3784 $p(y | \mathbf{w}) \propto p(y, \mathbf{w})$. The locality restriction in Equation 7.8 can be viewed as a conditional
 3785 independence assumption on the random variables y .

Algorithm 12 Generative process for the hidden Markov model

```

 $y_0 \leftarrow \diamond,$     $m \leftarrow 1$ 
repeat
     $y_m \sim \text{Categorical}(\lambda_{y_{m-1}})$             $\triangleright$  sample the current tag
     $w_m \sim \text{Categorical}(\phi_{y_m})$             $\triangleright$  sample the current word
until  $y_m = \blacklozenge$             $\triangleright$  terminate when the stop symbol is generated

```

3786 Naïve Bayes was introduced as a **generative model** — a probabilistic story that ex-
 3787 plains the observed data as well as the hidden label. A similar story can be constructed
 3788 for probabilistic sequence labeling: first, the tags are drawn from a prior distribution; next,
 3789 the tokens are drawn from a conditional likelihood. However, for inference to be tractable,
 3790 additional independence assumptions are required. First, the probability of each token
 3791 depends only on its tag, and not on any other element in the sequence:

$$p(w | y) = \prod_{m=1}^M p(w_m | y_m). \quad [7.32]$$

3792 Second, each tag y_m depends only on its predecessor,

$$p(y) = \prod_{m=1}^M p(y_m | y_{m-1}), \quad [7.33]$$

3793 where $y_0 = \diamond$ in all cases. Due to this **Markov assumption**, probabilistic sequence labeling
 3794 models are known as **hidden Markov models** (HMMs).

3795 The generative process for the hidden Markov model is shown in Algorithm 12. Given
 3796 the parameters λ and ϕ , we can compute $p(w, y)$ for any token sequence w and tag se-
 3797 quence y . The HMM is often represented as a **graphical model** (Wainwright and Jordan,
 3798 2008), as shown in Figure 7.2. This representation makes the independence assumptions
 3799 explicit: if a variable v_1 is probabilistically conditioned on another variable v_2 , then there
 3800 is an arrow $v_2 \rightarrow v_1$ in the diagram. If there are no arrows between v_1 and v_2 , they
 3801 are **conditionally independent**, given each variable's **Markov blanket**. In the hidden
 3802 Markov model, the Markov blanket for each tag y_m includes the “parent” y_{m-1} , and the
 3803 “children” y_{m+1} and w_m .³

3804 It is important to reflect on the implications of the HMM independence assumptions.
 3805 A non-adjacent pair of tags y_m and y_n are conditionally independent; if $m < n$ and we
 3806 are given y_{n-1} , then y_m offers no additional information about y_n . However, if we are
 3807 not given any information about the tags in a sequence, then all tags are probabilistically
 3808 coupled.

³In general graphical models, a variable's Markov blanket includes its parents, children, and its children's other parents (Murphy, 2012).

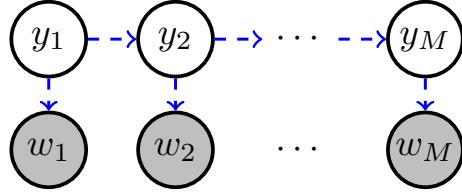


Figure 7.2: Graphical representation of the hidden Markov model. Arrows indicate probabilistic dependencies.

3809 7.4.1 Estimation

3810 The hidden Markov model has two groups of parameters:

3811 **Emission probabilities.** The probability $p_e(w_m | y_m; \phi)$ is the emission probability, since
3812 the words are treated as probabilistically “emitted”, conditioned on the tags.

3813 **Transition probabilities.** The probability $p_t(y_m | y_{m-1}; \lambda)$ is the transition probability,
3814 since it assigns probability to each possible tag-to-tag transition.

Both of these groups of parameters are typically computed from smoothed relative frequency estimation on a labeled corpus (see § 6.2 for a review of smoothing). The unsmoothed probabilities are,

$$\begin{aligned}\phi_{k,i} &\triangleq \Pr(W_m = i | Y_m = k) = \frac{\text{count}(W_m = i, Y_m = k)}{\text{count}(Y_m = k)} \\ \lambda_{k,k'} &\triangleq \Pr(Y_m = k' | Y_{m-1} = k) = \frac{\text{count}(Y_m = k', Y_{m-1} = k)}{\text{count}(Y_{m-1} = k)}.\end{aligned}$$

3815 Smoothing is more important for the emission probability than the transition probability,
3816 because the vocabulary is much larger than the number of tags.

3817 7.4.2 Inference

3818 The goal of inference in the hidden Markov model is to find the highest probability tag
3819 sequence,

$$\hat{y} = \underset{y}{\operatorname{argmax}} p(y | w). \quad [7.34]$$

3820 As in Naïve Bayes, it is equivalent to find the tag sequence with the highest *log*-probability,
3821 since the logarithm is a monotonically increasing function. It is furthermore equivalent
3822 to maximize the joint probability $p(y, w) = p(y | w) \times p(w) \propto p(y | w)$, which is pro-
3823 portional to the conditional probability. Putting these observations together, the inference

problem can be reformulated as,

$$\hat{\mathbf{y}} = \operatorname{argmax}_{\mathbf{y}} \log p(\mathbf{y}, \mathbf{w}). \quad [7.35]$$

We can now apply the HMM independence assumptions:

$$\log p(\mathbf{y}, \mathbf{w}) = \log p(\mathbf{y}) + \log p(\mathbf{w} \mid \mathbf{y}) \quad [7.36]$$

$$= \sum_{m=1}^{M+1} \log p_Y(y_m \mid y_{m-1}) + \log p_{W|Y}(w_m \mid y_m) \quad [7.37]$$

$$= \sum_{m=1}^{M+1} \log \lambda_{y_m, y_{m-1}} + \log \phi_{y_m, w_m} \quad [7.38]$$

$$= \sum_{m=1}^{M+1} s_m(y_m, y_{m-1}), \quad [7.39]$$

where,

$$s_m(y_m, y_{m-1}) \triangleq \log \lambda_{y_m, y_{m-1}} + \log \phi_{y_m, w_m}, \quad [7.40]$$

and,

$$\phi_{\diamond, w} = \begin{cases} 1, & w = \blacksquare \\ 0, & \text{otherwise,} \end{cases} \quad [7.41]$$

which ensures that the stop tag \diamond can only be applied to the final token \blacksquare .

This derivation shows that HMM inference can be viewed as an application of the Viterbi decoding algorithm, given an appropriately defined scoring function. The local score $s_m(y_m, y_{m-1})$ can be interpreted probabilistically,

$$s_m(y_m, y_{m-1}) = \log p_y(y_m \mid y_{m-1}) + \log p_{w|y}(w_m \mid y_m) \quad [7.42]$$

$$= \log p(y_m, w_m \mid y_{m-1}). \quad [7.43]$$

Now recall the definition of the Viterbi variables,

$$v_m(y_m) = \max_{y_{m-1}} s_m(y_m, y_{m-1}) + v_{m-1}(y_{m-1}) \quad [7.44]$$

$$= \max_{y_{m-1}} \log p(y_m, w_m \mid y_{m-1}) + v_{m-1}(y_{m-1}). \quad [7.45]$$

By setting $v_{m-1}(y_{m-1}) = \max_{\mathbf{y}_{1:m-2}} \log p(\mathbf{y}_{1:m-1}, \mathbf{w}_{1:m-1})$, we obtain the recurrence,

$$v_m(y_m) = \max_{y_{m-1}} \log p(y_m, w_m \mid y_{m-1}) + \max_{\mathbf{y}_{1:m-2}} \log p(\mathbf{y}_{1:m-1}, \mathbf{w}_{1:m-1}) \quad [7.46]$$

$$= \max_{\mathbf{y}_{1:m-1}} \log p(y_m, w_m \mid y_{m-1}) + \log p(\mathbf{y}_{1:m-1}, \mathbf{w}_{1:m-1}) \quad [7.47]$$

$$= \max_{\mathbf{y}_{1:m-1}} \log p(\mathbf{y}_{1:m}, \mathbf{w}_{1:m}). \quad [7.48]$$

In words, the Viterbi variable $v_m(y_m)$ is the log probability of the best tag sequence ending in y_m , joint with the word sequence $w_{1:m}$. The log probability of the best complete tag sequence is therefore,

$$\max_{\mathbf{y}_{1:M}} \log p(\mathbf{y}_{1:M+1}, \mathbf{w}_{1:M+1}) = v_{M+1}(\spadesuit) \quad [7.49]$$

***Viterbi as an example of the max-product algorithm** The Viterbi algorithm can also be implemented using probabilities, rather than log-probabilities. In this case, each $v_m(y_m)$ is equal to,

$$v_m(y_m) = \max_{\mathbf{y}_{1:m-1}} p(\mathbf{y}_{1:m-1}, y_m, \mathbf{w}_{1:m}) \quad [7.50]$$

$$= \max_{y_{m-1}} p(y_m, w_m | y_{m-1}) \times \max_{\mathbf{y}_{1:m-2}} p(\mathbf{y}_{1:m-2}, y_{m-1}, \mathbf{w}_{1:m-1}) \quad [7.51]$$

$$= \max_{y_{m-1}} p(y_m, w_m | y_{m-1}) \times v_{m-1}(y_{m-1}) \quad [7.52]$$

$$= p_{w|y}(w_m | y_m) \times \max_{y_{m-1}} p_y(y_m | y_{m-1}) \times v_{m-1}(y_{m-1}). \quad [7.53]$$

3827 Each Viterbi variable is computed by *maximizing* over a set of *products*. Thus, the Viterbi
 3828 algorithm is a special case of the **max-product algorithm** for inference in graphical mod-
 3829 els (Wainwright and Jordan, 2008). However, the product of probabilities tends towards
 3830 zero over long sequences, so the log-probability version of Viterbi is recommended in
 3831 practical implementations.

3832 7.5 Discriminative sequence labeling with features

3833 Today, hidden Markov models are rarely used for supervised sequence labeling. This is
 3834 because HMMs are limited to only two phenomena:

- 3835 • word-tag compatibility, via the emission probability $p_{W|Y}(w_m | y_m)$;
- 3836 • local context, via the transition probability $p_Y(y_m | y_{m-1})$.

3837 The Viterbi algorithm permits the inclusion of richer information in the local scoring func-
 3838 tion $\psi(\mathbf{w}_{1:M}, y_m, y_{m-1}, m)$, which can be defined as a weighted sum of arbitrary local *fea-*
 3839 *tures*,

$$\psi(\mathbf{w}, y_m, y_{m-1}, m) = \boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{w}, y_m, y_{m-1}, m), \quad [7.54]$$

3840 where \mathbf{f} is a locally-defined feature function, and $\boldsymbol{\theta}$ is a vector of weights.

The local decomposition of the scoring function Ψ is reflected in a corresponding decomposition of the feature function:

$$\Psi(\mathbf{w}, \mathbf{y}) = \sum_{m=1}^{M+1} \psi(\mathbf{w}, y_m, y_{m-1}, m) \quad [7.55]$$

$$= \theta \cdot \mathbf{f}(\mathbf{w}, y_m, y_{m-1}, m) \quad [7.56]$$

$$= \theta \cdot \sum_{m=1}^{M+1} \mathbf{f}(\mathbf{w}, y_m, y_{m-1}, m) \quad [7.57]$$

$$= \theta \cdot \mathbf{f}^{(\text{global})}(\mathbf{w}, \mathbf{y}_{1:M}), \quad [7.58]$$

3841 where $\mathbf{f}^{(\text{global})}(\mathbf{w}, \mathbf{y})$ is a global feature vector, which is a sum of local feature vectors,

$$\mathbf{f}^{(\text{global})}(\mathbf{w}, \mathbf{y}) = \sum_{m=1}^{M+1} \mathbf{f}(\mathbf{w}_{1:M}, y_m, y_{m-1}, m), \quad [7.59]$$

3842 with $y_{M+1} = \diamond$ and $y_0 = \diamond$ by construction.

3843 Let's now consider what additional information these features might encode.

3844 **Word affix features.** Consider the problem of part-of-speech tagging on the first four
3845 lines of the poem *Jabberwocky* (Carroll, 1917):

3846 (7.3) 'Twas brillig, and the slithy toves
3847 Did gyre and gimble in the wabe:
3848 All mimsy were the borogoves,
3849 And the mome raths outgrabe.

3850 Many of these words were made up by the author of the poem, so a corpus would offer
3851 no information about their probabilities of being associated with any particular part of
3852 speech. Yet it is not so hard to see what their grammatical roles might be in this passage.
3853 Context helps: for example, the word *slithy* follows the determiner *the*, so it is probably a
3854 noun or adjective. Which do you think is more likely? The suffix *-thy* is found in a number
3855 of adjectives, like *frothy*, *healthy*, *pithy*, *worthy*. It is also found in a handful of nouns — e.g.,
3856 *apathy*, *sympathy* — but nearly all of these have the longer coda *-pathy*, unlike *slithy*. So the
3857 suffix gives some evidence that *slithy* is an adjective, and indeed it is: later in the text we
3858 find that it is a combination of the adjectives *lithe* and *slimy*.⁴

⁴Morphology is the study of how words are formed from smaller linguistic units. chapter 9 touches on computational approaches to morphological analysis. See Bender (2013) for an overview of the underlying linguistic principles, and Haspelmath and Sims (2013) or Lieber (2015) for a full treatment.

3859 **Fine-grained context.** The hidden Markov model captures contextual information in the
 3860 form of part-of-speech tag bigrams. But sometimes, the necessary contextual information
 3861 is more specific. Consider the noun phrases *this fish* and *these fish*. Many part-of-speech
 3862 tagsets distinguish between singular and plural nouns, but do not distinguish between
 3863 singular and plural determiners; for example, the well known **Penn Treebank** tagset fol-
 3864 lows these conventions. A hidden Markov model would be unable to correctly label *fish* as
 3865 singular or plural in both of these cases, because it only has access to two features: the pre-
 3866 ceding tag (determiner in both cases) and the word (*fish* in both cases). The classification-
 3867 based tagger discussed in § 7.1 had the ability to use preceding and succeeding words as
 3868 features, and it can also be incorporated into a Viterbi-based sequence labeler as a local
 3869 feature.

Example. Consider the tagging D J N (determiner, adjective, noun) for the sequence *the slithy toves*, so that

$$\begin{aligned} \mathbf{w} &= \text{the slithy toves} \\ \mathbf{y} &= \text{D J N}. \end{aligned}$$

Let's create the feature vector for this example, assuming that we have word-tag features (indicated by W), tag-tag features (indicated by T), and suffix features (indicated by M). You can assume that you have access to a method for extracting the suffix *-thy* from *slithy*, *-es* from *toves*, and \emptyset from *the*, indicating that this word has no suffix.⁵ The resulting feature vector is,

$$\begin{aligned} \mathbf{f}(\text{the slithy toves, D J N}) &= \mathbf{f}(\text{the slithy toves, D}, \diamond, 1) \\ &\quad + \mathbf{f}(\text{the slithy toves, J}, \text{D}, 2) \\ &\quad + \mathbf{f}(\text{the slithy toves, N}, \text{J}, 3) \\ &\quad + \mathbf{f}(\text{the slithy toves}, \blacklozenge, \text{N}, 4) \\ &= \{(T : \diamond, \text{D}), (W : \text{the}, \text{D}), (M : \emptyset, \text{D}), \\ &\quad (T : \text{D}, \text{J}), (W : \text{slithy}, \text{J}), (M : \text{-thy}, \text{J}), \\ &\quad (T : \text{J}, \text{N}), (W : \text{toves}, \text{N}), (M : \text{-es}, \text{N}) \\ &\quad (T : \text{N}, \blacklozenge)\}. \end{aligned}$$

3870 These examples show that local features can incorporate information that lies beyond
 3871 the scope of a hidden Markov model. Because the features are local, it is possible to apply
 3872 the Viterbi algorithm to identify the optimal sequence of tags. The remaining question

⁵Such a system is called a **morphological segmenter**. The task of morphological segmentation is briefly described in § 9.1.4; a well known segmenter is MORFESSOR (Creutz and Lagus, 2007). In real applications, a typical approach is to include features for all orthographic suffixes up to some maximum number of characters: for *slithy*, we would have suffix features for *-y*, *-hy*, and *-thy*.

3873 is how to estimate the weights on these features. § 2.3 presented three main types of
 3874 discriminative classifiers: perceptron, support vector machine, and logistic regression.
 3875 Each of these classifiers has a structured equivalent, enabling it to be trained from labeled
 3876 sequences rather than individual tokens.

3877 **7.5.1 Structured perceptron**

The perceptron classifier is trained by increasing the weights for features that are associated with the correct label, and decreasing the weights for features that are associated with incorrectly predicted labels:

$$\hat{y} = \operatorname{argmax}_{y \in \mathcal{Y}} \theta \cdot f(\mathbf{x}, y) \quad [7.60]$$

$$\theta^{(t+1)} \leftarrow \theta^{(t)} + f(\mathbf{x}, y) - f(\mathbf{x}, \hat{y}). \quad [7.61]$$

We can apply exactly the same update in the case of structure prediction,

$$\hat{\mathbf{y}} = \operatorname{argmax}_{\mathbf{y} \in \mathcal{Y}(\mathbf{w})} \theta \cdot f(\mathbf{w}, \mathbf{y}) \quad [7.62]$$

$$\theta^{(t+1)} \leftarrow \theta^{(t)} + f(\mathbf{w}, \mathbf{y}) - f(\mathbf{w}, \hat{\mathbf{y}}). \quad [7.63]$$

3878 This learning algorithm is called **structured perceptron**, because it learns to predict the
 3879 structured output \mathbf{y} . The only difference is that instead of computing \hat{y} by enumerating
 3880 the entire set \mathcal{Y} , the Viterbi algorithm is used to efficiently search the set of possible tag-
 3881 gings, \mathcal{Y}^M . Structured perceptron can be applied to other structured outputs as long as
 3882 efficient inference is possible. As in perceptron classification, weight averaging is crucial
 3883 to get good performance (see § 2.3.2).

Example For the example *they can fish*, suppose that the reference tag sequence is $\mathbf{y}^{(i)} =$
 N V V, but the tagger incorrectly returns the tag sequence $\hat{\mathbf{y}} = \text{N V N}$. Assuming a model
 with features for emissions (w_m, y_m) and transitions (y_{m-1}, y_m) , the corresponding structured
 perceptron update is:

$$\theta_{(fish,V)} \leftarrow \theta_{(fish,V)} + 1, \quad \theta_{(fish,N)} \leftarrow \theta_{(fish,N)} - 1 \quad [7.64]$$

$$\theta_{(V,V)} \leftarrow \theta_{(V,V)} + 1, \quad \theta_{(V,N)} \leftarrow \theta_{(V,N)} - 1 \quad [7.65]$$

$$\theta_{(V,\blacklozenge)} \leftarrow \theta_{(V,\blacklozenge)} + 1, \quad \theta_{(N,\blacklozenge)} \leftarrow \theta_{(N,\blacklozenge)} - 1. \quad [7.66]$$

3884 **7.5.2 Structured support vector machines**

3885 Large-margin classifiers such as the support vector machine improve on the perceptron by
 3886 pushing the classification boundary away from the training instances. The same idea can

3887 be applied to sequence labeling. A support vector machine in which the output is a struc-
 3888 tured object, such as a sequence, is called a **structured support vector machine** (Tsochan-
 3889 taridis et al., 2004).⁶

3890 In classification, we formalized the large-margin constraint as,

$$\forall \mathbf{y} \neq \mathbf{y}^{(i)}, \boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{x}, \mathbf{y}^{(i)}) - \boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{x}, \mathbf{y}) \geq 1, \quad [7.67]$$

3891 requiring a margin of at least 1 between the scores for all labels \mathbf{y} that are not equal to the
 3892 correct label $\mathbf{y}^{(i)}$. The weights $\boldsymbol{\theta}$ are then learned by constrained optimization (see § 2.4.2).

3893 This idea can be applied to sequence labeling by formulating an equivalent set of con-
 3894 straints for all possible labelings $\mathcal{Y}(\mathbf{w})$ for an input \mathbf{w} . However, there are two problems.
 3895 First, in sequence labeling, some predictions are more wrong than others: we may miss
 3896 only one tag out of fifty, or we may get all fifty wrong. We would like our learning algo-
 3897 rithm to be sensitive to this difference. Second, the number of constraints is equal to the
 3898 number of possible labelings, which is exponentially large in the length of the sequence.

3899 The first problem can be addressed by adjusting the constraint to require larger mar-
 3900 gins for more serious errors. Let $c(\mathbf{y}^{(i)}, \hat{\mathbf{y}}) \geq 0$ represent the *cost* of predicting label $\hat{\mathbf{y}}$ when
 3901 the true label is $\mathbf{y}^{(i)}$. We can then generalize the margin constraint,

$$\forall \mathbf{y}, \boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{w}^{(i)}, \mathbf{y}^{(i)}) - \boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{w}^{(i)}, \mathbf{y}) \geq c(\mathbf{y}^{(i)}, \mathbf{y}). \quad [7.68]$$

3902 This cost-augmented margin constraint specializes to the constraint in Equation 7.67 if we
 3903 choose the delta function $c(\mathbf{y}^{(i)}, \mathbf{y}) = \delta((\mathbf{y}^{(i)} \neq \mathbf{y}))$. A more expressive cost function is
 3904 the **Hamming cost**,

$$c(\mathbf{y}^{(i)}, \mathbf{y}) = \sum_{m=1}^M \delta(y_m^{(i)} \neq y_m), \quad [7.69]$$

3905 which computes the number of errors in \mathbf{y} . By incorporating the cost function as the
 3906 margin constraint, we require that the true labeling be separated from the alternatives by
 3907 a margin that is proportional to the number of incorrect tags in each alternative labeling.

The second problem is that the number of constraints is exponential in the length
 of the sequence. This can be addressed by focusing on the prediction $\hat{\mathbf{y}}$ that *maximally*
 violates the margin constraint. This prediction can be identified by solving the following
cost-augmented decoding problem:

$$\hat{\mathbf{y}} = \operatorname{argmax}_{\mathbf{y} \neq \mathbf{y}^{(i)}} \boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{w}^{(i)}, \mathbf{y}) - \boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{w}^{(i)}, \mathbf{y}^{(i)}) + c(\mathbf{y}^{(i)}, \mathbf{y}) \quad [7.70]$$

$$= \operatorname{argmax}_{\mathbf{y} \neq \mathbf{y}^{(i)}} \boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{w}^{(i)}, \mathbf{y}) + c(\mathbf{y}^{(i)}, \mathbf{y}), \quad [7.71]$$

⁶This model is also known as a **max-margin Markov network** (Taskar et al., 2003), emphasizing that the scoring function is constructed from a sum of components, which are Markov independent.

3908 where in the second line we drop the term $\theta \cdot f(\mathbf{w}^{(i)}, \mathbf{y}^{(i)})$, which is constant in \mathbf{y} .

We can now reformulate the margin constraint for sequence labeling,

$$\theta \cdot f(\mathbf{w}^{(i)}, \mathbf{y}^{(i)}) - \max_{\mathbf{y} \in \mathcal{Y}(\mathbf{w})} (\theta \cdot f(\mathbf{w}^{(i)}, \mathbf{y}) + c(\mathbf{y}^{(i)}, \mathbf{y})) \geq 0. \quad [7.72]$$

3909 If the score for $\theta \cdot f(\mathbf{w}^{(i)}, \mathbf{y}^{(i)})$ is greater than the cost-augmented score for all alternatives,
 3910 then the constraint will be met. The name “cost-augmented decoding” is due to the fact
 3911 that the objective includes the standard decoding problem, $\max_{\hat{\mathbf{y}} \in \mathcal{Y}(\mathbf{w})} \theta \cdot f(\mathbf{w}, \hat{\mathbf{y}})$, plus
 3912 an additional term for the cost. Essentially, we want to train against predictions that are
 3913 strong and wrong: they should score highly according to the model, yet incur a large loss
 3914 with respect to the ground truth. Training adjusts the weights to reduce the score of these
 3915 predictions.

3916 For cost-augmented decoding to be tractable, the cost function must decompose into
 3917 local parts, just as the feature function $f(\cdot)$ does. The Hamming cost, defined above,
 3918 obeys this property. To perform cost-augmented decoding using the Hamming cost, we
 3919 need only to add features $f_m(y_m) = \delta(y_m \neq y_m^{(i)})$, and assign a constant weight of 1 to
 3920 these features. Decoding can then be performed using the Viterbi algorithm.⁷

As with large-margin classifiers, it is possible to formulate the learning problem in an unconstrained form, by combining a regularization term on the weights and a Lagrangian for the constraints:

$$\min_{\theta} \frac{1}{2} \|\theta\|_2^2 - C \left(\sum_i \theta \cdot f(\mathbf{w}^{(i)}, \mathbf{y}^{(i)}) - \max_{\mathbf{y} \in \mathcal{Y}(\mathbf{w}^{(i)})} [\theta \cdot f(\mathbf{w}^{(i)}, \mathbf{y}) + c(\mathbf{y}^{(i)}, \mathbf{y})] \right), \quad [7.73]$$

3921 In this formulation, C is a parameter that controls the tradeoff between the regularization
 3922 term and the margin constraints. A number of optimization algorithms have been
 3923 proposed for structured support vector machines, some of which are discussed in § 2.4.2.
 3924 An empirical comparison by Kummerfeld et al. (2015) shows that stochastic subgradient
 3925 descent — which is essentially a cost-augmented version of the structured perceptron —
 3926 is highly competitive.

3927 7.5.3 Conditional random fields

3928 The **conditional random field** (CRF; Lafferty et al., 2001) is a conditional probabilistic
 3929 model for sequence labeling; just as structured perceptron is built on the perceptron clas-
 3930 sifier, conditional random fields are built on the logistic regression classifier.⁸ The basic

⁷Are there cost functions that do not decompose into local parts? Suppose we want to assign a constant loss c to any prediction $\hat{\mathbf{y}}$ in which k or more predicted tags are incorrect, and zero loss otherwise. This loss function is combinatorial over the predictions, and thus we cannot decompose it into parts.

⁸The name “conditional random field” is derived from **Markov random fields**, a general class of models in which the probability of a configuration of variables is proportional to a product of scores across pairs (or

3931 probability model is,

$$p(\mathbf{y} \mid \mathbf{w}) = \frac{\exp(\Psi(\mathbf{w}, \mathbf{y}))}{\sum_{\mathbf{y}' \in \mathcal{Y}(\mathbf{w})} \exp(\Psi(\mathbf{w}, \mathbf{y}'))}. \quad [7.74]$$

3932 This is almost identical to logistic regression (§ 2.5), but because the label space is now
 3933 sequences of tags, we require efficient algorithms for both **decoding** (searching for the
 3934 best tag sequence given a sequence of words \mathbf{w} and a model θ) and for **normalization**
 3935 (summing over all tag sequences). These algorithms will be based on the usual locality
 3936 assumption on the scoring function, $\Psi(\mathbf{w}, \mathbf{y}) = \sum_{m=1}^{M+1} \psi(\mathbf{w}, y_m, y_{m-1}, m)$.

3937 Decoding in CRFs

Decoding — finding the tag sequence $\hat{\mathbf{y}}$ that maximizes $p(\mathbf{y} \mid \mathbf{w})$ — is a direct application of the Viterbi algorithm. The key observation is that the decoding problem does not depend on the denominator of $p(\mathbf{y} \mid \mathbf{w})$,

$$\begin{aligned} \hat{\mathbf{y}} &= \operatorname{argmax}_{\mathbf{y}} \log p(\mathbf{y} \mid \mathbf{w}) \\ &= \operatorname{argmax}_{\mathbf{y}} \Psi(\mathbf{y}, \mathbf{w}) - \log \sum_{\mathbf{y}' \in \mathcal{Y}(\mathbf{w})} \exp \Psi(\mathbf{y}', \mathbf{w}) \\ &= \operatorname{argmax}_{\mathbf{y}} \Psi(\mathbf{y}, \mathbf{w}) = \operatorname{argmax}_{\mathbf{y}} \sum_{m=1}^{M+1} s_m(y_m, y_{m-1}). \end{aligned}$$

3938 This is identical to the decoding problem for structured perceptron, so the same Viterbi
 3939 recurrence as defined in Equation 7.22 can be used.

3940 Learning in CRFs

As with logistic regression, the weights θ are learned by minimizing the regularized negative log-probability,

$$\ell = \frac{\lambda}{2} \|\theta\|^2 - \sum_{i=1}^N \log p(\mathbf{y}^{(i)} \mid \mathbf{w}^{(i)}; \theta) \quad [7.75]$$

$$= \frac{\lambda}{2} \|\theta\|^2 - \sum_{i=1}^N \theta \cdot f(\mathbf{w}^{(i)}, \mathbf{y}^{(i)}) + \log \sum_{\mathbf{y}' \in \mathcal{Y}(\mathbf{w}^{(i)})} \exp (\theta \cdot f(\mathbf{w}^{(i)}, \mathbf{y}')), \quad [7.76]$$

more generally, cliques) of variables in a **factor graph**. In sequence labeling, the pairs of variables include all adjacent tags (y_m, y_{m-1}). The probability is *conditioned* on the words \mathbf{w} , which are always observed, motivating the term “conditional” in the name.

3941 where λ controls the amount of regularization. The final term in Equation 7.76 is a sum
 3942 over all possible labelings. This term is the log of the denominator in Equation 7.74, some-
 3943 times known as the **partition function**.⁹ There are $|\mathcal{Y}|^M$ possible labelings of an input of
 3944 size M , so we must again exploit the decomposition of the scoring function to compute
 3945 this sum efficiently.

The sum $\sum_{\mathbf{y} \in \mathcal{Y}^{w(i)}} \exp \Psi(\mathbf{y}, \mathbf{w})$ can be computed efficiently using the **forward recurrence**, which is closely related to the Viterbi recurrence. We first define a set of **forward variables**, $\alpha_m(y_m)$, which is equal to the sum of the scores of all paths leading to tag y_m at position m :

$$\alpha_m(y_m) \triangleq \sum_{\mathbf{y}_{1:m-1}} \exp \sum_{n=1}^m s_n(y_n, y_{n-1}) \quad [7.77]$$

$$= \sum_{\mathbf{y}_{1:m-1}} \prod_{n=1}^m \exp s_n(y_n, y_{n-1}). \quad [7.78]$$

Note the similarity to the definition of the Viterbi variable, $v_m(y_m) = \max_{\mathbf{y}_{1:m-1}} \sum_{n=1}^m s_n(y_n, y_{n-1})$. In the hidden Markov model, the Viterbi recurrence had an alternative interpretation as the max-product algorithm (see Equation 7.53); analogously, the forward recurrence is known as the **sum-product algorithm**, because of the form of [7.78]. The forward variable can also be computed through a recurrence:

$$\alpha_m(y_m) = \sum_{\mathbf{y}_{1:m-1}} \prod_{n=1}^m \exp s_n(y_n, y_{n-1}) \quad [7.79]$$

$$= \sum_{y_{m-1}} (\exp s_m(y_m, y_{m-1})) \sum_{\mathbf{y}_{1:m-2}} \prod_{n=1}^{m-1} \exp s_n(y_n, y_{n-1}) \quad [7.80]$$

$$= \sum_{y_{m-1}} (\exp s_m(y_m, y_{m-1})) \times \alpha_{m-1}(y_{m-1}). \quad [7.81]$$

Using the forward recurrence, it is possible to compute the denominator of the conditional probability,

$$\sum_{\mathbf{y} \in \mathcal{Y}(\mathbf{w})} \Psi(\mathbf{w}, \mathbf{y}) = \sum_{\mathbf{y}_{1:M}} s_{M+1}(\blacklozenge, y_M) \prod_{m=1}^M s_m(y_m, y_{m-1}) \quad [7.82]$$

$$= \alpha_{M+1}(\blacklozenge). \quad [7.83]$$

⁹The terminology of “potentials” and “partition functions” comes from statistical mechanics (Bishop, 2006).

The conditional log-likelihood can be rewritten,

$$\ell = \frac{\lambda}{2} \|\boldsymbol{\theta}\|^2 - \sum_{i=1}^N \boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{w}^{(i)}, \mathbf{y}^{(i)}) + \log \alpha_{M+1}(\blacklozenge). \quad [7.84]$$

3946 Probabilistic programming environments, such as TORCH (Collobert et al., 2011) and
 3947 DYNET (Neubig et al., 2017), can compute the gradient of this objective using automatic
 3948 differentiation. The programmer need only implement the forward algorithm as a com-
 3949 putation graph.

As in logistic regression, the gradient of the likelihood with respect to the parameters is a difference between observed and expected feature counts:

$$\frac{d\ell}{d\theta_j} = \lambda\theta_j + \sum_{i=1}^N E[f_j(\mathbf{w}^{(i)}, \mathbf{y})] - f_j(\mathbf{w}^{(i)}, \mathbf{y}^{(i)}), \quad [7.85]$$

3950 where $f_j(\mathbf{w}^{(i)}, \mathbf{y}^{(i)})$ refers to the count of feature j for token sequence $\mathbf{w}^{(i)}$ and tag se-
 3951 quence $\mathbf{y}^{(i)}$. The expected feature counts are computed “under the hood” when automatic
 3952 differentiation is applied to Equation 7.84 (Eisner, 2016).

3953 Before the widespread use of automatic differentiation, it was common to compute
 3954 the feature expectations from marginal tag probabilities $p(y_m | \mathbf{w})$. These marginal prob-
 3955 abilities are sometimes useful on their own, and can be computed using the **forward-**
 3956 **backward algorithm**. This algorithm combines the forward recurrence with an equivalent
 3957 **backward recurrence**, which traverses the input from w_M back to w_1 .

3958 *Forward-backward algorithm

Marginal probabilities over tag bigrams can be written as,¹⁰

$$\Pr(Y_{m-1} = k', Y_m = k | \mathbf{w}) = \frac{\sum_{\mathbf{y}: Y_m=k, Y_{m-1}=k'} \prod_{n=1}^M \exp s_n(y_n, y_{n-1})}{\sum_{\mathbf{y}'} \prod_{n=1}^M \exp s_n(y'_n, y'_{n-1})}. \quad [7.86]$$

The numerator sums over all tag sequences that include the transition $(Y_{m-1} = k') \rightarrow (Y_m = k)$. Because we are only interested in sequences that include the tag bigram, this sum can be decomposed into three parts: the *prefixes* $\mathbf{y}_{1:m-1}$, terminating in $Y_{m-1} = k'$; the

¹⁰Recall the notational convention of upper-case letters for random variables, e.g. Y_m , and lower case letters for specific values, e.g., y_m , so that $Y_m = k$ is interpreted as the event of random variable Y_m taking the value k .

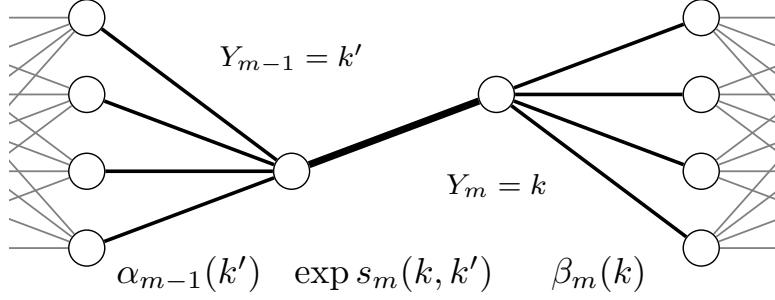


Figure 7.3: A schematic illustration of the computation of the marginal probability $\Pr(Y_{m-1} = k', Y_m = k)$, using the forward score $\alpha_{m-1}(k')$ and the backward score $\beta_m(k)$.

transition $(Y_{m-1} = k') \rightarrow (Y_m = k)$; and the *suffixes* $\mathbf{y}_{m:M}$, beginning with the tag $Y_m = k$:

$$\sum_{\mathbf{y}: Y_m = k, Y_{m-1} = k'} \prod_{n=1}^M \exp s_n(y_n, y_{n-1}) = \sum_{\mathbf{y}_{1:m-1}: Y_{m-1} = k'} \prod_{n=1}^{m-1} \exp s_n(y_n, y_{n-1}) \times \exp s_m(k, k') \times \sum_{\mathbf{y}_{m:M}: Y_m = k} \prod_{n=m+1}^{M+1} \exp s_n(y_n, y_{n-1}). \quad [7.87]$$

The result is product of three terms: a score that sums over all the ways to get to the position $(Y_{m-1} = k')$, a score for the transition from k' to k , and a score that sums over all the ways of finishing the sequence from $(Y_m = k)$. The first term of Equation 7.87 is equal to the **forward variable**, $\alpha_{m-1}(k')$. The third term — the sum over ways to finish the sequence — can also be defined recursively, this time moving over the trellis from right to left, which is known as the **backward recurrence**:

$$\beta_m(k) \triangleq \sum_{\mathbf{y}_{m:M}: Y_m = k} \prod_{n=m}^{M+1} \exp s_n(y_n, y_{n-1}) \quad [7.88]$$

$$= \sum_{k' \in \mathcal{Y}} \exp s_{m+1}(k', k) \sum_{\mathbf{y}_{m+1:M}: Y_m = k'} \prod_{n=m+1}^{M+1} \exp s_n(y_n, y_{n-1}) \quad [7.89]$$

$$= \sum_{k' \in \mathcal{Y}} \exp s_{m+1}(k', k) \times \beta_{m+1}(k'). \quad [7.90]$$

³⁹⁵⁹ To understand this computation, compare with the forward recurrence in Equation 7.81.

In practice, numerical stability demands that we work in the log domain,

$$\log \alpha_m(k) = \log \sum_{k' \in \mathcal{Y}} \exp (\log s_m(k, k') + \log \alpha_{m-1}(k')) \quad [7.91]$$

$$\log \beta_{m-1}(k) = \log \sum_{k' \in \mathcal{Y}} \exp (\log s_m(k', k) + \log \beta_m(k')). \quad [7.92]$$

3960 The application of the forward and backward probabilities is shown in Figure 7.3.
 3961 Both the forward and backward recurrences operate on the trellis, which implies a space
 3962 complexity $\mathcal{O}(MK)$. Because both recurrences require computing a sum over K terms at
 3963 each node in the trellis, their time complexity is $\mathcal{O}(MK^2)$.

3964 7.6 Neural sequence labeling

3965 In neural network approaches to sequence labeling, we construct a vector representa-
 3966 tion for each tagging decision, based on the word and its context. Neural networks can
 3967 perform tagging as a per-token classification decision, or they can be combined with the
 3968 Viterbi algorithm to tag the entire sequence globally.

3969 7.6.1 Recurrent neural networks

Recurrent neural networks (RNNs) were introduced in chapter 6 as a language modeling technique, in which the context at token m is summarized by a recurrently-updated vector,

$$\mathbf{h}_m = g(\mathbf{x}_m, \mathbf{h}_{m-1}), \quad m = 1, 2, \dots, M,$$

3970 where \mathbf{x}_m is the vector **embedding** of the token w_m and the function g defines the recur-
 3971 rence. The starting condition \mathbf{h}_0 is an additional parameter of the model. The long short-
 3972 term memory (LSTM) is a more complex recurrence, in which a memory cell is through a
 3973 series of gates, avoiding repeated application of the non-linearity. Despite these bells and
 3974 whistles, both models share the basic architecture of recurrent updates across a sequence,
 3975 and both will be referred to as RNNs here.

A straightforward application of RNNs to sequence labeling is to score each tag y_m as a linear function of \mathbf{h}_m :

$$\psi_m(y) = \beta_y \cdot \mathbf{h}_m \quad [7.93]$$

$$\hat{y}_m = \underset{y}{\operatorname{argmax}} \psi_m(y). \quad [7.94]$$

3976 The score $\psi_m(y)$ can also be converted into a probability distribution using the usual soft-
 3977 max operation,

$$p(y | \mathbf{w}_{1:m}) = \frac{\exp \psi_m(y)}{\sum_{y' \in \mathcal{Y}} \exp \psi_m(y')}. \quad [7.95]$$

3978 Using this transformation, it is possible to train the tagger from the negative log-likelihood
 3979 of the tags, as in a conditional random field. Alternatively, a hinge loss or margin loss
 3980 objective can be constructed from the raw scores $\psi_m(y)$.

The hidden state \mathbf{h}_m accounts for information in the input leading up to position m , but it ignores the subsequent tokens, which may also be relevant to the tag y_m . This can be addressed by adding a second RNN, in which the input is reversed, running the recurrence from w_M to w_1 . This is known as a **bidirectional recurrent neural network** (Graves and Schmidhuber, 2005), and is specified as:

$$\overleftarrow{\mathbf{h}}_m = g(\mathbf{x}_m, \overleftarrow{\mathbf{h}}_{m+1}), \quad m = 1, 2, \dots, M. \quad [7.96]$$

3981 The hidden states of the left-to-right RNN are denoted $\overrightarrow{\mathbf{h}}_m$. The left-to-right and right-to-
 3982 left vectors are concatenated, $\mathbf{h}_m = [\overleftarrow{\mathbf{h}}_m; \overrightarrow{\mathbf{h}}_m]$. The scoring function in Equation 7.93 is
 3983 applied to this concatenated vector.

3984 Bidirectional RNN tagging has several attractive properties. Ideally, the representa-
 3985 tion \mathbf{h}_m summarizes the useful information from the surrounding context, so that it is not
 3986 necessary to design explicit features to capture this information. If the vector \mathbf{h}_m is an ad-
 3987 equate summary of this context, then it may not even be necessary to perform the tagging
 3988 jointly: in general, the gains offered by joint tagging of the entire sequence are diminished
 3989 as the individual tagging model becomes more powerful. Using backpropagation, the
 3990 word vectors \mathbf{x} can be trained “end-to-end”, so that they capture word properties that are
 3991 useful for the tagging task. Alternatively, if limited labeled data is available, we can use
 3992 word embeddings that are “pre-trained” from unlabeled data, using a language modeling
 3993 objective (as in § 6.3) or a related word embedding technique (see chapter 14). It is even
 3994 possible to combine both fine-tuned and pre-trained embeddings in a single model.

3995 **Neural structure prediction** The bidirectional recurrent neural network incorporates in-
 3996 formation from throughout the input, but each tagging decision is made independently.
 3997 In some sequence labeling applications, there are very strong dependencies between tags:
 3998 it may even be impossible for one tag to follow another. In such scenarios, the tagging
 3999 decision must be made jointly across the entire sequence.

4000 Neural sequence labeling can be combined with the Viterbi algorithm by defining the
 4001 local scores as:

$$s_m(y_m, y_{m-1}) = \beta_{y_m} \cdot \mathbf{h}_m + \eta_{y_{m-1}, y_m}, \quad [7.97]$$

4002 where \mathbf{h}_m is the RNN hidden state, β_{y_m} is a vector associated with tag y_m , and η_{y_{m-1}, y_m}
 4003 is a scalar parameter for the tag transition (y_{m-1}, y_m) . These local scores can then be
 4004 incorporated into the Viterbi algorithm for inference, and into the forward algorithm for
 4005 training. This model is shown in Figure 7.4. It can be trained from the conditional log-
 4006 likelihood objective defined in Equation 7.76, backpropagating to the tagging parameters

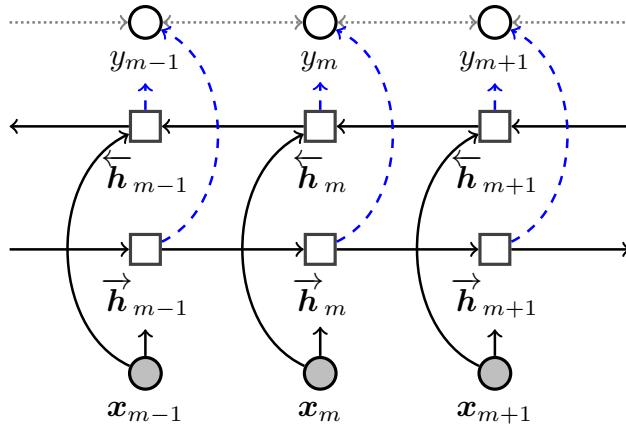


Figure 7.4: **Bidirectional LSTM** for sequence labeling. The solid lines indicate computation, the dashed lines indicate probabilistic dependency, and the dotted lines indicate the optional additional probabilistic dependencies between labels in the biLSTM-CRF.

4007 β and η , as well as the parameters of the RNN. This model is called the **LSTM-CRF**, due
 4008 to its combination of aspects of the long short-term memory and conditional random field
 4009 models (Huang et al., 2015).

4010 The LSTM-CRF is especially effective on the task of **named entity recognition** (Lample
 4011 et al., 2016), a sequence labeling task that is described in detail in § 8.3. This task has strong
 4012 dependencies between adjacent tags, so structure prediction is especially important.

4013 7.6.2 Character-level models

4014 As in language modeling, rare and unseen words are a challenge: if we encounter a word
 4015 that was not in the training data, then there is no obvious choice for the word embed-
 4016 ding x_m . One solution is to use a generic **unseen word** embedding for all such words.
 4017 However, in many cases, properties of unseen words can be guessed from their spellings.
 4018 For example, *whimsical* does not appear in the Universal Dependencies (UD) English Tree-
 4019 bank, yet the suffix *-al* makes it likely to be adjective; by the same logic, *unflinchingly* is
 4020 likely to be an adverb, and *barnacle* is likely to be a noun.

4021 In feature-based models, these morphological properties were handled by suffix fea-
 4022 tures; in a neural network, they can be incorporated by constructing the embeddings of
 4023 unseen words from their spellings or morphology. One way to do this is to incorporate
 4024 an additional layer of bidirectional RNNs, one for each word in the vocabulary (Ling
 4025 et al., 2015). For each such character-RNN, the inputs are the characters, and the output
 4026 is the concatenation of the final states of the left-facing and right-facing passes, $\phi_w =$

[$\vec{h}_{N_w}^{(w)}; \overleftarrow{h}_0^{(w)}$], where $\vec{h}_{N_w}^{(w)}$ is the final state of the right-facing pass for word w , and N_w is the number of characters in the word. The character RNN model is trained by back-propagation from the tagging objective. On the test data, the trained RNN is applied to out-of-vocabulary words (or all words), yielding inputs to the word-level tagging RNN. Other approaches to compositional word embeddings are described in § 14.7.1.

7.6.3 Convolutional Neural Networks for Sequence Labeling

One disadvantage of recurrent neural networks is that the architecture requires iterating through the sequence of inputs and predictions: each hidden vector h_m must be computed from the previous hidden vector h_{m-1} , before predicting the tag y_m . These iterative computations are difficult to parallelize, and fail to exploit the speedups offered by **graphics processing units (GPUs)** on operations such as matrix multiplication. **Convolutional neural networks** achieve better computational performance by predicting each label y_m from a set of matrix operations on the neighboring word embeddings, $x_{m-k:m+k}$ (Collobert et al., 2011). Because there is no hidden state to update, the predictions for each y_m can be computed in parallel. For more on convolutional neural networks, see § 3.4. Character-based word embeddings can also be computed using convolutional neural networks (Santos and Zadrozny, 2014).

7.7 *Unsupervised sequence labeling

In unsupervised sequence labeling, the goal is to induce a hidden Markov model from a corpus of *unannotated* text ($w^{(1)}, w^{(2)}, \dots, w^{(N)}$), where each $w^{(i)}$ is a sequence of length $M^{(i)}$. This is an example of the general problem of **structure induction**, which is the unsupervised version of structure prediction. The tags that result from unsupervised sequence labeling might be useful for some downstream task, or they might help us to better understand the language’s inherent structure. For part-of-speech tagging, it is common to use a tag dictionary that lists the allowed tags for each word, simplifying the problem (Christodoulopoulos et al., 2010).

Unsupervised learning in hidden Markov models can be performed using the **Baum-Welch algorithm**, which combines the forward-backward algorithm (§ 7.5.3) with expectation-maximization (EM; § 5.1.2). In the M-step, the HMM parameters from expected counts:

$$\Pr(W = i \mid Y = k) = \phi_{k,i} = \frac{E[\text{count}(W = i, Y = k)]}{E[\text{count}(Y = k)]}$$

$$\Pr(Y_m = k \mid Y_{m-1} = k') = \lambda_{k',k} = \frac{E[\text{count}(Y_m = k, Y_{m-1} = k')]}{E[\text{count}(Y_{m-1} = k')]} \quad 4052$$

4053 The expected counts are computed in the E-step, using the forward and backward
 4054 recurrences. The local scores follow the usual definition for hidden Markov models,

$$s_m(k, k') = \log p_E(w_m | Y_m = k; \phi) + \log p_T(Y_m = k | Y_{m-1} = k'; \lambda). \quad [7.98]$$

The expected transition counts for a single instance are,

$$E[\text{count}(Y_m = k, Y_{m-1} = k') | \mathbf{w}] = \sum_{m=1}^M \Pr(Y_{m-1} = k', Y_m = k | \mathbf{w}) \quad [7.99]$$

$$= \frac{\sum_{\mathbf{y}: Y_m=k, Y_{m-1}=k'} \prod_{n=1}^M \exp s_n(y_n, y_{n-1})}{\sum_{\mathbf{y}'} \prod_{n=1}^M \exp s_n(y'_n, y'_{n-1})}. \quad [7.100]$$

As described in § 7.5.3, these marginal probabilities can be computed from the forward-backward recurrence,

$$\Pr(Y_{m-1} = k', Y_m = k | \mathbf{w}) = \frac{\alpha_{m-1}(k') \times \exp s_m(k, k') \times \beta_m(k)}{\alpha_{M+1}(\blacklozenge)}. \quad [7.101]$$

In a hidden Markov model, each element of the forward-backward computation has a special interpretation:

$$\alpha_{m-1}(k') = p(Y_{m-1} = k', \mathbf{w}_{1:m-1}) \quad [7.102]$$

$$\exp s_m(k, k') = p(Y_m = k, w_m | Y_{m-1} = k') \quad [7.103]$$

$$\beta_m(k) = p(\mathbf{w}_{m+1:M} | Y_m = k). \quad [7.104]$$

Applying the conditional independence assumptions of the hidden Markov model (defined in Algorithm 12), the product is equal to the joint probability of the tag bigram and the entire input,

$$\begin{aligned} \alpha_{m-1}(k') \times \exp s_m(k, k') \times \beta_m(k) &= p(Y_{m-1} = k', \mathbf{w}_{1:m-1}) \\ &\quad \times p(Y_m = k, w_m | Y_{m-1} = k') \\ &\quad \times p(\mathbf{w}_{m+1:M} | Y_m = k) \\ &= p(Y_{m-1} = k', Y_m = k, \mathbf{w}_{1:M}). \end{aligned} \quad [7.105]$$

Dividing by $\alpha_{M+1}(\blacklozenge) = p(\mathbf{w}_{1:M})$ gives the desired probability,

$$\frac{\alpha_{m-1}(k') \times s_m(k, k') \times \beta_m(k)}{\alpha_{M+1}(\blacklozenge)} = \frac{p(Y_{m-1} = k', Y_m = k, \mathbf{w}_{1:M})}{p(\mathbf{w}_{1:M})} \quad [7.106]$$

$$= \Pr(Y_{m-1} = k', Y_m = k | \mathbf{w}_{1:M}). \quad [7.107]$$

4055 The expected emission counts can be computed in a similar manner, using the product
 4056 $\alpha_m(k) \times \beta_m(k)$.

4057 **7.7.1 Linear dynamical systems**

4058 The forward-backward algorithm can be viewed as Bayesian state estimation in a discrete
 4059 state space. In a continuous state space, $\mathbf{y}_m \in \mathbb{R}^K$, the equivalent algorithm is the **Kalman**
 4060 **smoother**. It also computes marginals $p(\mathbf{y}_m | \mathbf{x}_{1:M})$, using a similar two-step algorithm
 4061 of forward and backward passes. Instead of computing a trellis of values at each step, the
 4062 Kalman smoother computes a probability density function $q_{\mathbf{y}_m}(\mathbf{y}_m; \boldsymbol{\mu}_m, \Sigma_m)$, character-
 4063 ized by a mean $\boldsymbol{\mu}_m$ and a covariance Σ_m around the latent state. Connections between the
 4064 Kalman smoother and the forward-backward algorithm are elucidated by Minka (1999)
 4065 and Murphy (2012).

4066 **7.7.2 Alternative unsupervised learning methods**

As noted in § 5.5, expectation-maximization is just one of many techniques for structure induction. One alternative is to use **Markov Chain Monte Carlo (MCMC)** sampling algorithms, which are briefly described in § 5.5.1. For the specific case of sequence labeling, Gibbs sampling can be applied by iteratively sampling each tag y_m conditioned on all the others (Finkel et al., 2005):

$$p(y_m | \mathbf{y}_{-m}, \mathbf{w}_{1:M}) \propto p(w_m | y_m) p(y_m | \mathbf{y}_{-m}). \quad [7.108]$$

4067 Gibbs Sampling has been applied to unsupervised part-of-speech tagging by Goldwater
 4068 and Griffiths (2007). **Beam sampling** is a more sophisticated sampling algorithm, which
 4069 randomly draws entire sequences $\mathbf{y}_{1:M}$, rather than individual tags y_m ; this algorithm
 4070 was applied to unsupervised part-of-speech tagging by Van Gael et al. (2009). Spectral
 4071 learning (see § 5.5.2) can also be applied to sequence labeling. By factoring matrices of
 4072 co-occurrence counts of word bigrams and trigrams (Song et al., 2010; Hsu et al., 2012), it
 4073 is possible to obtain globally optimal estimates of the transition and emission parameters,
 4074 under mild assumptions.

4075 **7.7.3 Semiring notation and the generalized viterbi algorithm**

The Viterbi and Forward recurrences can each be performed over probabilities or log probabilities, yielding a total of four closely related recurrences. These four recurrence scan in fact be expressed as a single recurrence in a more general notation, known as **semiring algebra**. Let the symbols \oplus and \otimes represent generalized addition and multiplication respectively.¹¹ Given these operators, a generalized Viterbi recurrence is denoted,

$$v_m(k) = \bigoplus_{k' \in \mathcal{Y}} s_m(k, k') \otimes v_{m-1}(k'). \quad [7.109]$$

¹¹In a semiring, the addition and multiplication operators must both obey associativity, and multiplication must distribute across addition; the addition operator must be commutative; there must be additive and multiplicative identities $\bar{0}$ and $\bar{1}$, such that $a \oplus \bar{0} = a$ and $a \otimes \bar{1} = a$; and there must be a multiplicative annihilator $\bar{0}$, such that $a \otimes \bar{0} = \bar{0}$.

4076 Each recurrence that we have seen so far is a special case of this generalized Viterbi
 4077 recurrence:

- 4078 • In the max-product Viterbi recurrence over probabilities, the \oplus operation corre-
 4079 sponds to maximization, and the \otimes operation corresponds to multiplication.
- 4080 • In the forward recurrence over probabilities, the \oplus operation corresponds to addition,
 4081 and the \otimes operation corresponds to multiplication.
- 4082 • In the max-product Viterbi recurrence over log-probabilities, the \oplus operation corre-
 4083 sponds to maximization, and the \otimes operation corresponds to addition.¹²
- 4084 • In the forward recurrence over log-probabilities, the \oplus operation corresponds to log-
 4085 addition, $a \oplus b = \log(e^a + e^b)$. The \otimes operation corresponds to addition.

4086 The mathematical abstraction offered by semiring notation can be applied to the soft-
 4087 ware implementations of these algorithms, yielding concise and modular implemen-
 4088 tations. For example, in the OPENFST library, generic operations are parametrized by the
 4089 choice of semiring (Allauzen et al., 2007).

4090 Exercises

- 4091 1. Extend the example in § 7.3.1 to the sentence *they can can fish*, meaning that “they can
 4092 put fish into cans.” Build the trellis for this example using the weights in Table 7.1,
 4093 and identify the best-scoring tag sequence. If the scores for noun and verb are tied,
 4094 then you may assume that the backpointer always goes to noun.
- 4095 2. Using the tagset $\mathcal{Y} = \{N, V\}$, and the feature set $f(\mathbf{w}, y_m, y_{m-1}, m) = \{(w_m, y_m), (y_m, y_{m-1})\}$,
 4096 show that there is no set of weights that give the correct tagging for both *they can*
 4097 *fish* (N V V) and *they can can fish* (N V V N).
- 4098 3. Work out what happens if you train a structured perceptron on the two exam-
 4099 ples mentioned in the previous problem, using the transition and emission features
 4100 (y_m, y_{m-1}) and (y_m, w_m) . Initialize all weights at 0, and assume that the Viterbi algo-
 4101 rithm always chooses *N* when the scores for the two tags are tied, so that the initial
 4102 prediction for *they can fish* is N N N.
- 4103 4. Consider the garden path sentence, *The old man the boat*. Given word-tag and tag-tag
 4104 features, what inequality in the weights must hold for the correct tag sequence to
 4105 outscore the garden path tag sequence for this example?

¹²This is sometimes called the **tropical semiring**, in honor of the Brazilian mathematician Imre Simon.

- 4106 5. Using the weights in Table 7.1, explicitly compute the log-probabilities for all pos-
 4107 sible taggings of the input *fish can*. Verify that the forward algorithm recovers the
 4108 aggregate log probability.
- 4109 6. Sketch out an algorithm for a variant of Viterbi that returns the top-*n* label se-
 4110 quences. What is the time and space complexity of this algorithm?
- 4111 7. Show how to compute the marginal probability $\Pr(y_{m-2} = k, y_m = k' \mid \mathbf{w}_{1:M})$, in
 4112 terms of the forward and backward variables, and the potentials $s_n(y_n, y_{n-1})$.
- 4113 8. Suppose you receive a stream of text, where some of tokens have been replaced at
 4114 random with *NOISE*. For example:
- 4115 • Source: *I try all things, I achieve what I can*
 - 4116 • Message received: *I try NOISE NOISE, I NOISE what I NOISE*
- 4117 Assume you have access to a pre-trained bigram language model, which gives prob-
 4118 abilities $p(w_m \mid w_{m-1})$. These probabilities can be assumed to be non-zero for all
 4119 bigrams.
- 4120 Show how to use the Viterbi algorithm to recover the source by maximizing the
 4121 bigram language model log-probability. Specifically, set the scores $s_m(y_m, y_{m-1})$ so
 4122 that the Viterbi algorithm selects a sequence of words that maximizes the bigram
 4123 language model log-probability, while leaving the non-noise tokens intact. Your
 4124 solution should not modify the logic of the Viterbi algorithm, it should only set the
 4125 scores $s_m(y_m, y_{m-1})$.
- 4126 9. Let $\alpha(\cdot)$ and $\beta(\cdot)$ indicate the forward and backward variables as defined in § 7.5.3.
 4127 Prove that $\alpha_{M+1}(\blacklozenge) = \beta_0(\lozenge) = \sum_y \alpha_m(y) \beta_m(y), \forall m \in \{1, 2, \dots, M\}$.
- 4128 10. Consider an RNN tagging model with a tanh activation function on the hidden
 4129 layer, and a hinge loss on the output. (The problem also works for the margin loss
 4130 and negative log-likelihood.) Suppose you initialize all parameters to zero: this in-
 4131 cludes the word embeddings that make up \mathbf{x} , the transition matrix Θ , the output
 4132 weights β , and the initial hidden state \mathbf{h}_0 .
 - 4133 a) Prove that for any data and for any gradient-based learning algorithm, all pa-
 4134 rameters will be stuck at zero.
 - 4135 b) Would a sigmoid activation function avoid this problem?

4136 Chapter 8

4137 Applications of sequence labeling

4138 Sequence labeling has applications throughout natural language processing. This chapter
4139 focuses on part-of-speech tagging, morpho-syntactic attribute tagging, named entity
4140 recognition, and tokenization. It also touches briefly on two applications to interactive
4141 settings: dialogue act recognition and the detection of code-switching points between
4142 languages.

4143 8.1 Part-of-speech tagging

4144 The **syntax** of a language is the set of principles under which sequences of words are
4145 judged to be grammatically acceptable by fluent speakers. One of the most basic syntactic
4146 concepts is the **part-of-speech** (POS), which refers to the syntactic role of each word in a
4147 sentence. This concept was used informally in the previous chapter, and you may have
4148 some intuitions from your own study of English. For example, in the sentence *We like*
4149 *vegetarian sandwiches*, you may already know that *we* and *sandwiches* are nouns, *like* is a
4150 verb, and *vegetarian* is an adjective. These labels depend on the context in which the word
4151 appears: in *she eats like a vegetarian*, the word *like* is a preposition, and the word *vegetarian*
4152 is a noun.

4153 Parts-of-speech can help to disentangle or explain various linguistic problems. Recall
4154 Chomsky's proposed distinction in chapter 6:

- 4155 (8.1) a. Colorless green ideas sleep furiously.
- 4156 b. * Ideas colorless furiously green sleep.

4157 One difference between these two examples is that the first contains part-of-speech transi-
4158 tions that are typical in English: adjective to adjective, adjective to noun, noun to verb,
4159 and verb to adverb. The second example contains transitions that are unusual: noun to
4160 adjective and adjective to verb. The ambiguity in a headline like,

4161 (8.2) Teacher Strikes Idle Children

4162 can also be explained in terms of parts of speech: in the interpretation that was likely
 4163 intended, *strikes* is a noun and *idle* is a verb; in the alternative explanation, *strikes* is a verb
 4164 and *idle* is an adjective.

4165 Part-of-speech tagging is often taken as a early step in a natural language processing
 4166 pipeline. Indeed, parts-of-speech provide features that can be useful for many of the
 4167 tasks that we will encounter later, such as parsing (chapter 10), coreference resolution
 4168 (chapter 15), and relation extraction (chapter 17).

4169 **8.1.1 Parts-of-Speech**

4170 The **Universal Dependencies** project (UD) is an effort to create syntactically-annotated
 4171 corpora across many languages, using a single annotation standard (Nivre et al., 2016). As
 4172 part of this effort, they have designed a part-of-speech **tagset**, which is meant to capture
 4173 word classes across as many languages as possible.¹ This section describes that inventory,
 4174 giving rough definitions for each of tags, along with supporting examples.

4175 Part-of-speech tags are **morphosyntactic**, rather than semantic, categories. This means
 4176 that they describe words in terms of how they pattern together and how they are inter-
 4177 nally constructed (e.g., what suffixes and prefixes they include). For example, you may
 4178 think of a noun as referring to objects or concepts, and verbs as referring to actions or
 4179 events. But events can also be nouns:

4180 (8.3) ... the **howling** of the **shrieking** storm.

4181 Here *howling* and *shrieking* are events, but grammatically they act as a noun and adjective
 4182 respectively.

4183 **The Universal Dependency part-of-speech tagset**

4184 The UD tagset is broken up into three groups: open class tags, closed class tags, and
 4185 “others.”

4186 **Open class tags** Nearly all languages contain nouns, verbs, adjectives, and adverbs.²
 4187 These are all **open word classes**, because new words can easily be added to them. The
 4188 UD tagset includes two other tags that are open classes: proper nouns and interjections.

4189 • **Nouns** (UD tag: NOUN) tend to describe entities and concepts, e.g.,

¹The UD tagset builds on earlier work from Petrov et al. (2012), in which a set of twelve universal tags was identified by creating mappings from tagsets for individual languages.

²One prominent exception is Korean, which some linguists argue does not have adjectives Kim (2002).

4190 (8.4) **Toes** are scarce among veteran **blubber men**.

4191 In English, nouns tend to follow determiners and adjectives, and can play the subject
4192 role in the sentence. They can be marked for the plural number by an *-s* suffix.

- 4193 • **Proper nouns** (PROPN) are tokens in names, which uniquely specify a given entity,

4194 (8.5) “**Moby Dick?**” shouted **Ahab**.

- 4195 • **Verbs** (VERB), according to the UD guidelines, “typically signal events and ac-
4196 tions.” But they are also defined grammatically: they “can constitute a minimal
4197 predicate in a clause, and govern the number and types of other constituents which
4198 may occur in a clause.”³

4199 (8.6) “**Moby Dick?**” shouted Ahab.

4200 (8.7) Shall we **keep chasing** this murderous fish?

4201 English verbs tend to come in between the subject and some number of direct ob-
4202 jects, depending on the verb. They can be marked for **tense** and **aspect** using suffixes
4203 such as *-ed* and *-ing*. (These suffixes are an example of **inflectional morphology**,
4204 which is discussed in more detail in § 9.1.4.)

- 4205 • **Adjectives** (ADJ) describe properties of entities,

4206 (8.8) a. Shall we keep chasing this **murderous** fish?

4207 b. Toes are **scarce** among **veteran** blubber men.

4208 In the second example, *scarce* is a predicative adjective, linked to the subject by the
4209 **copula verb** *are*. In contrast, *murderous* and *veteran* are attributive adjectives, modi-
4210 fying the noun phrase in which they are embedded.

- 4211 • **Adverbs** (ADV) describe properties of events, and may also modify adjectives or
4212 other adverbs:

4213 (8.9) a. It is not down on any map; true places **never** are.

4214 b. ...**treacherously** hidden beneath the loveliest tints of azure

4215 c. Not drowned **entirely**, though.

- 4216 • **Interjections** (INTJ) are used in exclamations, e.g.,

4217 (8.10) **Aye aye!** it was that accursed white whale that razed me.

³<http://universaldependencies.org/u/pos/VERB.html>

4218 **Closed class tags** Closed word classes rarely receive new members. They are sometimes
 4219 referred to as **function words** — as opposed to **content words** — as they have little lexical
 4220 meaning of their own, but rather, help to organize the components of the sentence.

- 4221 • **Adpositions** (ADP) describe the relationship between a complement (usually a noun
 4222 phrase) and another unit in the sentence, typically a noun or verb phrase.

- 4223 (8.11) a. Toes are scarce **among** veteran blubber men.
 4224 b. It is not **down on** any map.
 4225 c. Give not thyself **up** then.

4226 As the examples show, English generally uses prepositions, which are adpositions
 4227 that appear before their complement. (An exception is *ago*, as in, *we met three days*
 4228 *ago*). Postpositions are used in other languages, such as Japanese and Turkish.

- 4229 • **Auxiliary verbs** (AUX) are a closed class of verbs that add information such as
 4230 tense, aspect, person, and number.

- 4231 (8.12) a. **Shall** we keep chasing this murderous fish?
 4232 b. What the white whale was to Ahab, **has been** hinted.
 4233 c. Ahab **must** use tools.
 4234 d. Meditation and water **are** wedded forever.
 4235 e. Toes **are** scarce among veteran blubber men.

4236 The final example is a copula verb, which is also tagged as an auxiliary in the UD
 4237 corpus.

- 4238 • **Coordinating conjunctions** (CCONJ) express relationships between two words or
 4239 phrases, which play a parallel role:

- 4240 (8.13) Meditation **and** water are wedded forever.

- 4241 • **Subordinating conjunctions** (SCONJ) link two clauses, making one syntactically
 4242 subordinate to the other:

- 4243 (8.14) It is the easiest thing in the world for a man to look as **if** he had a great
 4244 secret in him.

4245 Note that

- 4246 • **Pronouns** (PRON) are words that substitute for nouns or noun phrases.

- 4247 (8.15) a. Be **it what it will**, I'll go to **it** laughing.

4248 b. I try all things, I achieve **what** I can.

4249 The example includes the personal pronouns *I* and *it*, as well as the relative pronoun
4250 *what*. Other pronouns include *myself*, *somebody*, and *nothing*.

- 4251 • **Determiners** (DET) provide additional information about the nouns or noun phrases
4252 that they modify:

- 4253 (8.16) a. What **the** white whale was to Ahab, has been hinted.
4254 b. It is not down on **any** map.
4255 c. I try **all** things ...
4256 d. Shall we keep chasing **this** murderous fish?

4257 Determiners include articles (*the*), possessive determiners (*their*), demonstratives
4258 (*this murderous fish*), and quantifiers (*any map*).

- 4259 • **Numerals** (NUM) are an infinite but closed class, which includes integers, fractions,
4260 and decimals, regardless of whether spelled out or written in numerical form.

- 4261 (8.17) a. How then can this **one** small heart beat.
4262 b. I am going to put him down for the **three hundredth**.

- 4263 • **Particles** (PART) are a catch-all of function words that combine with other words or
4264 phrases, but do not meet the conditions of the other tags. In English, this includes
4265 the infinitival *to*, the possessive marker, and negation.

- 4266 (8.18) a. Better **to** sleep with a sober cannibal than a drunk Christian.
4267 b. So man's insanity is heaven's sense
4268 c. It is **not** down on any map

4269 As the second example shows, the possessive marker is not considered part of the
4270 same token as the word that it modifies, so that *man's* is split into two tokens. (Tok-
4271 enization is described in more detail in § 8.4.) A non-English example of a particle
4272 is the Japanese question marker *ka*:⁴

- 4273 (8.19) Sensei desu ka
4274 Teacher is ?
 Is she a teacher?

⁴In this notation, the first line is the transliterated Japanese text, the second line is a token-to-token **gloss**, and the third line is the translation.

4275 **Other** The remaining UD tags include punctuation (PUN) and symbols (SYM). Punc-
 4276 tuation is purely structural — e.g., commas, periods, colons — while symbols can carry
 4277 content of their own. Examples of symbols include dollar and percentage symbols, math-
 4278 ematical operators, emoticons, emojis, and internet addresses. A final catch-all tag is X,
 4279 which is used for words that cannot be assigned another part-of-speech category. The X
 4280 tag is also used in cases of **code switching** (between languages), described in § 8.5.

4281 **Other tagsets**

4282 Prior to the Universal Dependency treebank, part-of-speech tagging was performed us-
 4283 ing language-specific tagsets. The dominant tagset for English was designed as part of
 4284 the **Penn Treebank** (PTB), and it includes 45 tags — more than three times as many as
 4285 the UD tagset. This granularity is reflected in distinctions between singular and plural
 4286 nouns, verb tenses and aspects, possessive and non-possessive pronouns, comparative
 4287 and superlative adjectives and adverbs (e.g., *faster, fastest*), and so on. The Brown corpus
 4288 includes a tagset that is even more detailed, with 87 tags (Francis, 1964), including special
 4289 tags for individual auxiliary verbs such as *be, do, and have*.

4290 Different languages make different distinctions, and so the PTB and Brown tagsets are
 4291 not appropriate for a language such as Chinese, which does not mark the verb tense (Xia,
 4292 2000); nor for Spanish, which marks every combination of person and number in the
 4293 verb ending; nor for German, which marks the case of each noun phrase. Each of these
 4294 languages requires more detail than English in some areas of the tagset, and less in other
 4295 areas. The strategy of the Universal Dependencies corpus is to design a coarse-grained
 4296 tagset to be used across all languages, and then to additionally annotate language-specific
 4297 **morphosyntactic attributes**, such as number, tense, and case. The attribute tagging task
 4298 is described in more detail in § 8.2.

4299 Social media such as Twitter have been shown to require tagsets of their own (Gimpel
 4300 et al., 2011). Such corpora contain some tokens that are not equivalent to anything en-
 4301 countered in a typical written corpus: e.g., emoticons, URLs, and hashtags. Social media
 4302 also includes dialectal words like *gonna* ('going to', e.g. *We gonna be fine*) and *Ima* ('I'm
 4303 going to', e.g., *Ima tell you one more time*), which can be analyzed either as non-standard
 4304 orthography (making tokenization impossible), or as lexical items in their own right. In
 4305 either case, it is clear that existing tags like NOUN and VERB cannot handle cases like *Ima*,
 4306 which combine aspects of the noun and verb. Gimpel et al. (2011) therefore propose a new
 4307 set of tags to deal with these cases.

4308 **8.1.2 Accurate part-of-speech tagging**

4309 Part-of-speech tagging is the problem of selecting the correct tag for each word in a sen-
 4310 tence. Success is typically measured by accuracy on an annotated test set, which is simply
 4311 the fraction of tokens that were tagged correctly.

4312 **Baselines**

4313 A simple baseline for part-of-speech tagging is to choose the most common tag for each
4314 word. For example, in the Universal Dependencies treebank, the word *talk* appears 96
4315 times, and 85 of those times it is labeled as a VERB: therefore, this baseline will always
4316 predict VERB for this word. For words that do not appear in the training corpus, the base-
4317 line simply guesses the most common tag overall, which is NOUN. In the Penn Treebank,
4318 this simple baseline obtains accuracy above 92%. A more rigorous evaluation is the accu-
4319 racy on **out-of-vocabulary words**, which are not seen in the training data. Tagging these
4320 words correctly requires attention to the context and the word's internal structure.

4321 **Contemporary approaches**

4322 Conditional random fields and structured perceptron perform at or near the state-of-the-
4323 art for part-of-speech tagging in English. For example, (Collins, 2002) achieved 97.1%
4324 accuracy on the Penn Treebank, using a structured perceptron with the following base
4325 features (originally introduced by Ratnaparkhi (1996)):

- 4326 • current word, w_m
- 4327 • previous words, w_{m-1}, w_{m-2}
- 4328 • next words, w_{m+1}, w_{m+2}
- 4329 • previous tag, y_{m-1}
- 4330 • previous two tags, (y_{m-1}, y_{m-2})
- 4331 • for rare words:
 - 4332 – first k characters, up to $k = 4$
 - 4333 – last k characters, up to $k = 4$
 - 4334 – whether w_m contains a number, uppercase character, or hyphen.

4335 Similar results for the PTB data have been achieved using conditional random fields (CRFs;
4336 Toutanova et al., 2003).

4337 More recent work has demonstrated the power of neural sequence models, such as the
4338 **long short-term memory (LSTM)** (§ 7.6). Plank et al. (2016) apply a CRF and a bidirec-
4339 tional LSTM to twenty-two languages in the UD corpus, achieving an average accuracy
4340 of 94.3% for the CRF, and 96.5% with the bi-LSTM. Their neural model employs three
4341 types of embeddings: fine-tuned word embeddings, which are updated during training;
4342 pre-trained word embeddings, which are never updated, but which help to tag out-of-
4343 vocabulary words; and character-based embeddings. The character-based embeddings
4344 are computed by running an LSTM on the individual characters in each word, thereby
4345 capturing common orthographic patterns such as prefixes, suffixes, and capitalization.
4346 Extensive evaluations show that these additional embeddings are crucial to their model's
4347 success.

word	PTB tag	UD tag	UD attributes
<i>The</i>	DT	DET	DEFINITE=DEF PRONTYPE=ART
<i>German</i>	JJ	ADJ	DEGREE=POS
<i>Expressionist</i>	NN	NOUN	NUMBER=SING
<i>movement</i>	NN	NOUN	NUMBER=SING
<i>was</i>	VBD	AUX	MOOD=IND NUMBER=SING PERSON=3 TENSE=PAST VERBFORM=FIN
<i>destroyed</i>	VBN	VERB	TENSE=PAST VERBFORM=PART VOICE=PASS
<i>as</i>	IN	ADP	
<i>a</i>	DT	DET	DEFINITE=IND PRONTYPE=ART
<i>result</i>	NN	NOUN	NUMBER=SING
.	.	PUNCT	

Figure 8.1: UD and PTB part-of-speech tags, and UD morphosyntactic attributes. Example selected from the UD 1.4 English corpus.

4348 8.2 Morphosyntactic Attributes

4349 There is considerably more to say about a word than whether it is a noun or a verb: in En-
 4350 glish, verbs are distinguish by features such tense and aspect, nouns by number, adjectives
 4351 by degree, and so on. These features are language-specific: other languages distinguish
 4352 other features, such as **case** (the role of the noun with respect to the action of the sen-
 4353 tence, which is marked in languages such as Latin and German⁵) and **evidentiality** (the
 4354 source of information for the speaker’s statement, which is marked in languages such as
 4355 Turkish). In the UD corpora, these attributes are annotated as feature-value pairs for each
 4356 token.⁶

4357 An example is shown in Figure 8.1. The determiner *the* is marked with two attributes:
 4358 **PRONTYPE=ART**, which indicates that it is an **article** (as opposed to another type of deter-

⁵Case is marked in English for some personal pronouns, e.g., *She saw her, They saw them*.

⁶The annotation and tagging of morphosyntactic attributes can be traced back to earlier work on Turkish (Oflazer and Kuruöz, 1994) and Czech (Hajič and Hladká, 1998). MULTEXT-East was an early multilingual corpus to include morphosyntactic attributes (Dimitrova et al., 1998).

4359 miner or pronominal modifier), and DEFINITE=DEF, which indicates that it is a **definite**
4360 **article** (referring to a specific, known entity). The verbs are each marked with several
4361 attributes. The auxiliary verb *was* is third-person, singular, past tense, finite (conjugated),
4362 and indicative (describing an event that has happened or is currently happenings); the
4363 main verb *destroyed* is in participle form (so there is no additional person and number
4364 information), past tense, and passive voice. Some, but not all, of these distinctions are
4365 reflected in the PTB tags VBD (past-tense verb) and VBN (past participle).

4366 While there are thousands of papers on part-of-speech tagging, there is comparatively
4367 little work on automatically labeling morphosyntactic attributes. Faruqui et al. (2016)
4368 train a support vector machine classification model, using a minimal feature set that in-
4369 cludes the word itself, its prefixes and suffixes, and type-level information listing all pos-
4370 sible morphosyntactic attributes for each word and its neighbors. Mueller et al. (2013) use
4371 a conditional random field (CRF), in which the tag space consists of all observed com-
4372 binations of morphosyntactic attributes (e.g., the tag would be DEF+ART for the word
4373 *the* in Figure 8.1). This massive tag space is managed by decomposing the feature space
4374 over individual attributes, and pruning paths through the trellis. More recent work has
4375 employed bidirectional LSTM sequence models. For example, Pinter et al. (2017) train
4376 a bidirectional LSTM sequence model. The input layer and hidden vectors in the LSTM
4377 are shared across attributes, but each attribute has its own output layer, culminating in
4378 a softmax over all attribute values, e.g. $y_t^{\text{NUMBER}} \in \{\text{SING}, \text{PLURAL}, \dots\}$. They find that
4379 character-level information is crucial, especially when the amount of labeled data is lim-
4380 ited.

4381 Evaluation is performed by first computing recall and precision for each attribute.
4382 These scores can then be averaged at either the type or token level to obtain micro- or
4383 macro-*F*-MEASURE. Pinter et al. (2017) evaluate on 23 languages in the UD treebank,
4384 reporting a median micro-*F*-MEASURE of 0.95. Performance is strongly correlated with the
4385 size of the labeled dataset for each language, with a few outliers: for example, Chinese is
4386 particularly difficult, because although the dataset is relatively large (10^5 tokens in the UD
4387 1.4 corpus), only 6% of tokens have any attributes, offering few useful labeled instances.

4388 8.3 Named Entity Recognition

4389 A classical problem in information extraction is to recognize and extract mentions of
4390 **named entities** in text. In news documents, the core entity types are people, locations, and
4391 organizations; more recently, the task has been extended to include amounts of money,
4392 percentages, dates, and times. In item 8.20a (Figure 8.2), the named entities include: *The*
4393 *U.S. Army*, an organization; *Atlanta*, a location; and *May 14, 1864*, a date. Named en-
4394 tity recognition is also a key task in **biomedical natural language processing**, with entity
4395 types including proteins, DNA, RNA, and cell lines (e.g., Collier et al., 2000; Ohta et al.,
4396 2002). Figure 8.2 shows an example from the GENIA corpus of biomedical research ab-

- (8.20) a. *The U.S. Army captured Atlanta on May 14, 1864*
 B-ORG I-ORG I-ORG O B-LOC O B-DATE I-DATE I-DATE I-DATE
 b. *Number of glucocorticoid receptors in lymphocytes and ...*
 O O B-PROTEIN I-PROTEIN O B-CELLTYPE O ...

Figure 8.2: BIO notation for named entity recognition. Example (8.20b) is drawn from the GENIA corpus of biomedical documents (Ohta et al., 2002).

4397 stracts.

4398 A standard approach to tagging named entity spans is to use discriminative sequence
 4399 labeling methods such as conditional random fields. However, the named entity recogni-
 4400 tion (NER) task would seem to be fundamentally different from sequence labeling tasks
 4401 like part-of-speech tagging: rather than tagging each token, the goal is to recover *spans*
 4402 of tokens, such as *The United States Army*.

4403 This is accomplished by the **BIO notation**, shown in Figure 8.2. Each token at the
 4404 beginning of a name span is labeled with a B- prefix; each token within a name span is la-
 4405 beled with an I- prefix. These prefixes are followed by a tag for the entity type, e.g. B-LOC
 4406 for the beginning of a location, and I-PROTEIN for the inside of a protein name. Tokens
 4407 that are not parts of name spans are labeled as O. From this representation, the entity
 4408 name spans can be recovered unambiguously. This tagging scheme is also advantageous
 4409 for learning: tokens at the beginning of name spans may have different properties than
 4410 tokens within the name, and the learner can exploit this. This insight can be taken even
 4411 further, with special labels for the last tokens of a name span, and for unique tokens in
 4412 name spans, such as *Atlanta* in the example in Figure 8.2. This is called BILOU notation,
 4413 and it can yield improvements in supervised named entity recognition (Ratinov and Roth,
 4414 2009).

Feature-based sequence labeling Named entity recognition was one of the first applications of conditional random fields (McCallum and Li, 2003). The use of Viterbi decoding restricts the feature function $f(\mathbf{w}, \mathbf{y})$ to be a sum of local features, $\sum_m f(\mathbf{w}, y_m, y_{m-1}, m)$, so that each feature can consider only local adjacent tags. Typical features include tag transitions, word features for w_m and its neighbors, character-level features for prefixes and suffixes, and “word shape” features for capitalization and other orthographic properties. As an example, base features for the word *Army* in the example in (8.20a) include:

(CURR-WORD:*Army*, PREV-WORD:*U.S.*, NEXT-WORD:*captured*, PREFIX-1:*A-*,
 PREFIX-2:*Ar-*, SUFFIX-1:*-y*, SUFFIX-2:*-my*, SHAPE:*Xxxx*)

4415 Features can also be obtained from a **gazetteer**, which is a list of known entity names. For
 4416 example, the U.S. Social Security Administration provides a list of tens of thousands of

- (1) 日文 章魚 怎麼 說?
 Japanese octopus how say
 How to say octopus in Japanese?
- (2) 日 文章 魚 怎麼 說?
 Japan essay fish how say

Figure 8.3: An example of tokenization ambiguity in Chinese (Sproat et al., 1996)

4417 given names — more than could be observed in any annotated corpus. Tokens or spans
 4418 that match an entry in a gazetteer can receive special features; this provides a way to
 4419 incorporate hand-crafted resources such as name lists in a learning-driven framework.

4420 **Neural sequence labeling for NER** Current research has emphasized neural sequence
 4421 labeling, using similar LSTM models to those employed in part-of-speech tagging (Ham-
 4422 merton, 2003; Huang et al., 2015; Lample et al., 2016). The bidirectional LSTM-CRF (Fig-
 4423 ure 7.4 in § 7.6) does particularly well on this task, due to its ability to model tag-to-tag
 4424 dependencies. However, Strubell et al. (2017) show that **convolutional neural networks**
 4425 can be equally accurate, with significant improvement in speed due to the efficiency of
 4426 implementing ConvNets on **graphics processing units (GPUs)**. The key innovation in
 4427 this work was the use of **dilated convolution**, which is described in more detail in § 3.4.

4428 8.4 Tokenization

4429 A basic problem for text analysis, first discussed in § 4.3.1, is to break the text into a se-
 4430 quence of discrete tokens. For alphabetic languages such as English, deterministic scripts
 4431 usually suffice to achieve accurate tokenization. However, in logographic writing systems
 4432 such as Chinese script, words are typically composed of a small number of characters,
 4433 without intervening whitespace. The tokenization must be determined by the reader, with
 4434 the potential for occasional ambiguity, as shown in Figure 8.3. One approach is to match
 4435 character sequences against a known dictionary (e.g., Sproat et al., 1996), using additional
 4436 statistical information about word frequency. However, no dictionary is completely com-
 4437 prehensive, and dictionary-based approaches can struggle with such out-of-vocabulary
 4438 words.

4439 Chinese word segmentation has therefore been approached as a supervised sequence
 4440 labeling problem. Xue et al. (2003) train a logistic regression classifier to make indepen-
 4441 dent segmentation decisions while moving a sliding window across the document. A set
 4442 of rules is then used to convert these individual classification decisions into an overall to-
 4443 kenization of the input. However, these individual decisions may be globally suboptimal,
 4444 motivating a structure prediction approach. Peng et al. (2004) train a conditional random

4445 field to predict labels of START or NONSTART on each character. More recent work has
 4446 employed neural network architectures. For example, Chen et al. (2015) use an LSTM-
 4447 CRF architecture, as described in § 7.6: they construct a trellis, in which each tag is scored
 4448 according to the hidden state of an LSTM, and tag-tag transitions are scored according
 4449 to learned transition weights. The best-scoring segmentation is then computed by the
 4450 Viterbi algorithm.

4451 8.5 Code switching

4452 Multilingual speakers and writers do not restrict themselves to a single language. **Code**
4453 **switching** is the phenomenon of switching between languages in speech and text (Auer,
4454 2013; Poplack, 1980). Written code switching has become more common in online social
4455 media, as in the following extract from the website of Canadian President Justin Trudeau:⁷

- 4456 (8.21) *Although everything written on this site est disponible en anglais*
is available in English
4457 *and in French, my personal videos seront bilingues*
will be bilingual

4458 Accurately analyzing such texts requires first determining which languages are being
4459 used. Furthermore, quantitative analysis of code switching can provide insights on the
4460 languages themselves and their relative social positions.

Code switching can be viewed as a sequence labeling problem, where the goal is to label each token as a candidate switch point. In the example above, the words *est*, *and*, and *seront* would be labeled as switch points. Solorio and Liu (2008) detect English-Spanish switch points using a supervised classifier, with features that include the word, its part-of-speech in each language (according to a supervised part-of-speech tagger), and the probabilities of the word and part-of-speech in each language. Nguyen and Dogruöz (2013) apply a conditional random field to the problem of detecting code switching between Turkish and Dutch.

Code switching is a special case of the more general problem of word level language identification, which Barman et al. (2014) address in the context of trilingual code switching between Bengali, English, and Hindi. They further observe an even more challenging phenomenon: intra-word code switching, such as the use of English suffixes with Bengali roots. They therefore mark each token as either (1) belonging to one of the three languages; (2) a mix of multiple languages; (3) “universal” (e.g., symbols, numbers, emoticons); or (4) undefined.

⁷As quoted in <http://blogues.lapresse.ca/lagace/2008/09/08/justin-trudeau-really-parfait-bilingue/>, accessed August 21, 2017.

Speaker	Dialogue Act	Utterance
A	YES-NO-QUESTION	<i>So do you go college right now?</i>
A	ABANDONED	<i>Are yo-</i>
B	YES-ANSWER	<i>Yeah,</i>
B	STATEMENT	<i>It's my last year [laughter].</i>
A	DECLARATIVE-QUESTION	<i>You're a, so you're a senior now.</i>
B	YES-ANSWER	<i>Yeah,</i>
B	STATEMENT	<i>I'm working on my projects trying to graduate [laughter]</i>
A	APPRECIATION	<i>Oh, good for you.</i>
B	BACKCHANNEL	<i>Yeah.</i>

Figure 8.4: An example of dialogue act labeling (Stolcke et al., 2000)

4476 8.6 Dialogue acts

4477 The sequence labeling problems that we have discussed so far have been over sequences
 4478 of word tokens or characters (in the case of tokenization). However, sequence labeling
 4479 can also be performed over higher-level units, such as **utterances**. **Dialogue acts** are la-
 4480 bels over utterances in a dialogue, corresponding roughly to the speaker’s intention —
 4481 the utterance’s **illocutionary force** (Austin, 1962). For example, an utterance may state a
 4482 proposition (*it is not down on any map*), pose a question (*shall we keep chasing this murderous*
 4483 *fish?*), or provide a response (*aye aye!*). Stolcke et al. (2000) describe how a set of 42 dia-
 4484 logue acts were annotated for the 1,155 conversations in the Switchboard corpus (Godfrey
 4485 et al., 1992).⁸

4486 An example is shown in Figure 8.4. The annotation is performed over UTTERANCES,
 4487 with the possibility of multiple utterances per **conversational turn** (in cases such as inter-
 4488 ruptions, an utterance may split over multiple turns). Some utterances are clauses (e.g., *So*
 4489 *do you go to college right now?*), while others are single words (e.g., *yeah*). Stolcke et al. (2000)
 4490 report that hidden Markov models (HMMs) achieve 96% accuracy on supervised utter-
 4491 ance segmentation. The labels themselves reflect the conversational goals of the speaker:
 4492 the utterance *yeah* functions as an answer in response to the question *you’re a senior now*,
 4493 but in the final line of the excerpt, it is a **backchannel** (demonstrating comprehension).

4494 For task of dialogue act labeling, Stolcke et al. (2000) apply a hidden Markov model.
 4495 The probability $p(w_m | y_m)$ must generate the entire sequence of words in the utterance,
 4496 and it is modeled as a trigram language model (§ 6.1). Stolcke et al. (2000) also account
 4497 for acoustic features, which capture the **prosody** of each utterance — for example, tonal
 4498 and rhythmic properties of speech, which can be used to distinguish dialogue acts such

⁸Dialogue act modeling is not restricted to speech; it is relevant in any interactive conversation. For example, Jeong et al. (2009) annotate a more limited set of **speech acts** in a corpus of emails and online forums.

4499 as questions and answers. These features are handled with an additional emission distri-
 4500 bution, $p(a_m | y_m)$, which is modeled with a probabilistic decision tree (Murphy, 2012).
 4501 While acoustic features yield small improvements overall, they play an important role in
 4502 distinguish questions from statements, and agreements from backchannels.

4503 Recurrent neural architectures for dialogue act labeling have been proposed by Kalch-
 4504 brenner and Blunsom (2013) and Ji et al. (2016), with strong empirical results. Both models
 4505 are recurrent at the utterance level, so that each complete utterance updates a hidden state.
 4506 The recurrent-convolutional network of Kalchbrenner and Blunsom (2013) uses convolu-
 4507 tion to obtain a representation of each individual utterance, while Ji et al. (2016) use a
 4508 second level of recurrence, over individual words. This enables their method to also func-
 4509 tion as a language model, giving probabilities over sequences of words in a document.

4510 Exercises

4511 1. Using the Universal Dependencies part-of-speech tags, annotate the following sen-
 4512 tences. You may examine the UD tagging guidelines. Tokenization is shown with
 4513 whitespace. Don't forget about punctuation.

- 4514 (8.22) a. I try all things , I achieve what I can .
 4515 b. It was that accursed white whale that razed me .
 4516 c. Better to sleep with a sober cannibal , than a drunk Christian .
 4517 d. Be it what it will , I 'll go to it laughing .

4518 2. Select three short sentences from a recent news article, and annotate them for UD
 4519 part-of-speech tags. Ask a friend to annotate the same three sentences without look-
 4520 ing at your annotations. Compute the rate of agreement, using the Kappa metric
 4521 defined in § 4.5.2. Then work together to resolve any disagreements.

4522 3. Choose one of the following morphosyntactic attributes: MOOD, TENSE, VOICE. Re-
 4523 search the definition of this attribute on the universal dependencies website, <http://universaldependencies.org/u/feat/index.html>. Returning to the ex-
 4524 amples in the first exercise, annotate all verbs for your chosen attribute. It may be
 4525 helpful to consult examples from an English-language universal dependencies cor-
 4526 pus, available at [https://github.com/UniversalDependencies/UD_English-EWT/
 4527 tree/master](https://github.com/UniversalDependencies/UD_English-EWT/tree/master).

4529 4. Download a dataset annotated for universal dependencies, such as the English Tree-
 4530 bank at [https://github.com/UniversalDependencies/UD_English-EWT/
 4531 tree/master](https://github.com/UniversalDependencies/UD_English-EWT/tree/master). This corpus is already segmented into training, development, and
 4532 test data.

- 4533 a) First, train a logistic regression or SVM classifier using character suffixes: char-
 4534 acter n-grams up to length 4. Compute the recall, precision, and *F*-MEASURE
 4535 on the development data.
- 4536 b) Next, augment your classifier using the same character suffixes of the preced-
 4537 ing and succeeding tokens. Again, evaluate your classifier on heldout data.
- 4538 c) Optionally, train a Viterbi-based sequence labeling model, using a toolkit such
 4539 as CRFSuite (<http://www.chokkan.org/software/crfsuite/>) or your
 4540 own Viterbi implementation. This is more likely to be helpful for attributes
 4541 in which agreement is required between adjacent words. For example, many
 4542 Romance languages require gender and number agreement for determiners,
 4543 nouns, and adjectives.
- 4544 5. Provide BIO-style annotation of the named entities (person, place, organization,
 4545 date, or product) in the following expressions:
- 4546 (8.23) a. The third mate was Flask, a native of Tisbury, in Martha's Vineyard.
 4547 b. Its official Nintendo announced today that they Will release the Nin-
 4548 tendo 3DS in north America march 27 (Ritter et al., 2011).
 4549 c. Jessica Reif, a media analyst at Merrill Lynch & Co., said, "If they can
 4550 get up and running with exclusive programming within six months, it
 4551 doesn't set the venture back that far."⁹
- 4552 6. Run the examples above through the online version of a named entity recogni-
 4553 tion tagger, such as the Allen NLP system here: <http://demo.allennlp.org/named->
 4554 entity-recognition. Do the predicted tags match your annotations?
- 4555 7. Build a whitespace tokenizer for English:
- 4556 a) Using the NLTK library, download the complete text to the novel *Alice in Won-*
 4557 *derland* (Carroll, 1865). Hold out the final 1000 words as a test set.
- 4558 b) Label each alphanumeric character as a segmentation point, $y_m = 1$ if m is
 4559 the final character of a token. Label every other character as $y_m = 0$. Then
 4560 concatenate all the tokens in the training and test sets. Make sure that the num-
 4561 ber of labels $\{y_m\}_{m=1}^M$ is identical to the number of characters $\{c_m\}_{m=1}^M$ in your
 4562 concatenated datasets.
- 4563 c) Train a logistic regression classifier to predict y_m , using the surrounding char-
 4564 acters $c_{m-5:m+5}$ as features. After training the classifier, run it on the test set,
 4565 using the predicted segmentation points to re-tokenize the text.

⁹From the Message Understanding Conference (MUC-7) dataset (Chinchor and Robinson, 1997).

- 4566 d) Compute the per-character segmentation accuracy on the test set. You should
4567 be able to get at least 88% accuracy.
4568 e) Print out a sample of segmented text from the test set, e.g.

4569 Thereareno mice in the air , I ' m afraid , but y oumight cat
4570 chabat , and that ' s very like a mouse , youknow . But
4571 docatseat bats , I wonder ?'

- 4572 8. Perform the following extensions to your tokenizer in the previous problem.

- 4573 a) Train a conditional random field sequence labeler, by incorporating the tag
4574 bigrams (y_{m-1}, y_m) as additional features. You may use a structured predic-
4575 tion library such as CRFSuite, or you may want to implement Viterbi yourself.
4576 Compare the accuracy with your classification-based approach.
- 4577 b) Compute the token-level performance: treating the original tokenization as
4578 ground truth, compute the number of true positives (tokens that are in both
4579 the ground truth and predicted tokenization), false positives (tokens that are in
4580 the predicted tokenization but not the ground truth), and false negatives (to-
4581 kens that are in the ground truth but not the predicted tokenization). Compute
4582 the F-measure.
4583 Hint: to match predicted and ground truth tokens, add “anchors” for the start
4584 character of each token. The number of true positives is then the size of the
4585 intersection of the sets of predicted and ground truth tokens.
- 4586 c) Apply the same methodology in a more practical setting: tokenization of Chi-
4587 nese, which is written without whitespace. You can find annotated datasets at
4588 <http://alias-i.com/lingpipe/demos/tutorial/chineseTokens/read-me.html>.

4590 **Chapter 9**

4591 **Formal language theory**

4592 We have now seen methods for learning to label individual words, vectors of word counts,
4593 and sequences of words; we will soon proceed to more complex structural transfor-
4594 mations. Most of these techniques could apply to counts or sequences from any discrete vo-
4595 cabulary; there is nothing fundamentally linguistic about, say, a hidden Markov model.
4596 This raises a basic question that this text has not yet considered: what is a language?

4597 This chapter will take the perspective of **formal language theory**, in which a language
4598 is defined as a set of **strings**, each of which is a sequence of elements from a finite alphabet.
4599 For interesting languages, there are an infinite number of strings that are in the language,
4600 and an infinite number of strings that are not. For example:

- 4601 • the set of all even-length sequences from the alphabet $\{a, b\}$, e.g., $\{\emptyset, aa, ab, ba, bb, aaaa, aaab, \dots\}$;
- 4602 • the set of all sequences from the alphabet $\{a, b\}$ that contain *aaa* as a substring, e.g.,
4603 $\{aaa, aaaa, baaa, aaab, \dots\}$;
- 4604 • the set of all sequences of English words (drawn from a finite dictionary) that con-
4605 tain at least one verb (a finite subset of the dictionary);
- 4606 • the PYTHON programming language.

4607 Formal language theory defines classes of languages and their computational prop-
4608 erties. Of particular interest is the computational complexity of solving the **membership**
4609 **problem** — determining whether a string is in a language. The chapter will focus on
4610 three classes of formal languages: regular, context-free, and “mildly” context-sensitive
4611 languages.

4612 A key insight of 20th century linguistics is that formal language theory can be usefully
4613 applied to natural languages such as English, by designing formal languages that cap-
4614 ture as many properties of the natural language as possible. For many such formalisms, a
4615 useful linguistic analysis comes as a byproduct of solving the membership problem. The

4616 membership problem can be generalized to the problems of *scoring* strings for their ac-
 4617 ceptability (as in language modeling), and of **transducing** one string into another (as in
 4618 translation).

4619 9.1 Regular languages

4620 If you have written a **regular expression**, then you have defined a **regular language**: a
 4621 regular language is any language that can be defined by a regular expression. Formally, a
 4622 regular expression can include the following elements:

- 4623 • A **literal character** drawn from some finite alphabet Σ .
- 4624 • The **empty string** ϵ .
- 4625 • The concatenation of two regular expressions RS , where R and S are both regular
 4626 expressions. The resulting expression accepts any string that can be decomposed
 4627 $x = yz$, where y is accepted by R and z is accepted by S .
- 4628 • The alternation $R \mid S$, where R and S are both regular expressions. The resulting
 4629 expression accepts a string x if it is accepted by R or it is accepted by S .
- 4630 • The **Kleene star** R^* , which accepts any string x that can be decomposed into a se-
 4631 quence of strings which are all accepted by R .
- 4632 • Parenthesization ((R)), which is used to limit the scope of the concatenation, alterna-
 4633 tion, and Kleene star operators.

4634 Here are some example regular expressions:

- 4635 • The set of all even length strings on the alphabet $\{a, b\}$: $((aa)|(ab)|(ba)|(bb))^*$
- 4636 • The set of all sequences of the alphabet $\{a, b\}$ that contain aaa as a substring: $(a|b)^*aaa(a|b)^*$
- 4637 • The set of all sequences of English words that contain at least one verb: W^*VW^* ,
 4638 where W is an alternation between all words in the dictionary, and V is an alterna-
 4639 tion between all verbs ($V \subseteq W$).

4640 This list does not include a regular expression for the Python programming language,
 4641 because this language is not regular — there is no regular expression that can capture its
 4642 syntax. We will discuss why towards the end of this section.

4643 Regular languages are **closed** under union, intersection, and concatenation. This means
 4644 that if two languages L_1 and L_2 are regular, then so are the languages $L_1 \cup L_2$, $L_1 \cap L_2$,
 4645 and the language of strings that can be decomposed as $s = tu$, with $s \in L_1$ and $t \in L_2$.
 4646 Regular languages are also closed under negation: if L is regular, then so is the language
 4647 $\bar{L} = \{s \notin L\}$.

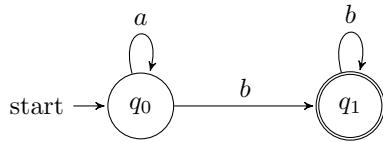


Figure 9.1: State diagram for the finite state acceptor M_1 .

9.1.1 Finite state acceptors

A regular expression defines a regular language, but does not give an algorithm for determining whether a string is in the language that it defines. **Finite state automata** are theoretical models of computation on regular languages, which involve transitions between a finite number of states. The most basic type of finite state automaton is the **finite state acceptor (FSA)**, which describes the computation involved in testing if a string is a member of a language. Formally, a finite state acceptor is a tuple $M = (Q, \Sigma, q_0, F, \delta)$, consisting of:

- a finite alphabet Σ of input symbols;
- a finite set of states $Q = \{q_0, q_1, \dots, q_n\}$;
- a start state $q_0 \in Q$;
- a set of final states $F \subseteq Q$;
- a transition function $\delta : Q \times (\Sigma \cup \{\epsilon\}) \rightarrow 2^Q$. The transition function maps from a state and an input symbol (or empty string ϵ) to a *set* of possible resulting states.

A **path** in M is a sequence of transitions, $\pi = t_1, t_2, \dots, t_N$, where each t_i traverses an arc in the transition function δ . The finite state acceptor M accepts a string ω if there is an accepting path, in which the initial transition t_1 begins at the start state q_0 , the final transition t_N terminates in a final state in Q , and the entire input ω is consumed.

Example

Consider the following FSA, M_1 .

$$\Sigma = \{a, b\} \quad [9.1]$$

$$Q = \{q_0, q_1\} \quad [9.2]$$

$$F = \{q_1\} \quad [9.3]$$

$$\delta = \{(q_0, a) \rightarrow q_0, (q_0, b) \rightarrow q_1, (q_1, b) \rightarrow q_1\}. \quad [9.4]$$

This FSA defines a language over an alphabet of two symbols, a and b . The transition function δ is written as a set of arcs: $(q_0, a) \rightarrow q_0$ says that if the machine is in state

4669 q_0 and reads symbol a , it stays in q_0 . Figure 9.1 provides a graphical representation of
 4670 M_1 . Because each pair of initial state and symbol has at most one resulting state, M_1 is
 4671 **deterministic**: each string ω induces at most one accepting path. Note that there are no
 4672 transitions for the symbol a in state q_1 ; if a is encountered in q_1 , then the acceptor is stuck,
 4673 and the input string is rejected.

4674 What strings does M_1 accept? The start state is q_0 , and we have to get to q_1 , since this
 4675 is the only final state. Any number of a symbols can be consumed in q_0 , but a b symbol is
 4676 required to transition to q_1 . Once there, any number of b symbols can be consumed, but
 4677 an a symbol cannot. So the regular expression corresponding to the language defined by
 4678 M_1 is a^*bb^* .

4679 Computational properties of finite state acceptors

4680 The key computational question for finite state acceptors is: how fast can we determine
 4681 whether a string is accepted? For deterministic FSAs, this computation can be performed
 4682 by Dijkstra's algorithm, with time complexity $\mathcal{O}(V \log V + E)$, where V is the number of
 4683 vertices in the FSA, and E is the number of edges (Cormen et al., 2009). Non-deterministic
 4684 FSAs (NFSAs) can include multiple transitions from a given symbol and state. Any NSFA
 4685 can be converted into a deterministic FSA, but the resulting automaton may have a num-
 4686 ber of states that is exponential in the number of size of the original NFSFA (Mohri et al.,
 4687 2002).

4688 9.1.2 Morphology as a regular language

4689 Many words have internal structure, such as prefixes and suffixes that shape their mean-
 4690 ing. The study of word-internal structure is the domain of **morphology**, of which there
 4691 are two main types:

- 4692 • **Derivational morphology** describes the use of affixes to convert a word from one
 4693 grammatical category to another (e.g., from the noun *grace* to the adjective *graceful*),
 4694 or to change the meaning of the word (e.g., from *grace* to *disgrace*).
- 4695 • **Inflectional morphology** describes the addition of details such as gender, number,
 4696 person, and tense (e.g., the *-ed* suffix for past tense in English).

4697 Morphology is a rich topic in linguistics, deserving of a course in its own right.¹ The
 4698 focus here will be on the use of finite state automata for morphological analysis. The

¹A good starting point would be a chapter from a linguistics textbook (e.g., Akmajian et al., 2010; Bender, 2013). A key simplification in this chapter is the focus on affixes at the sole method of derivation and inflection. English makes use of affixes, but also incorporates **apophony**, such as the inflection of *foot* to *feet*. Semitic languages like Arabic and Hebrew feature a template-based system of morphology, in which roots are triples of consonants (e.g., *ktb*), and words are created by adding vowels: *kataba* (Arabic: he wrote), *kutub* (books), *maktab* (desk). For more detail on morphology, see texts from Haspelmath and Sims (2013) and Lieber (2015).

4699 current section deals with derivational morphology; inflectional morphology is discussed
4700 in § 9.1.4.

4701 Suppose that we want to write a program that accepts only those words that are con-
4702 structed in accordance with the rules of English derivational morphology:

- 4703 (9.1) a. grace, graceful, gracefully, *gracelyful
4704 b. disgrace, *ungrace, disgraceful, disgracefully
4705 c. allure, *allureful, alluring, alluringly
4706 d. fairness, unfair, *disfair, fairly

4707 (Recall that the asterisk indicates that a linguistic example is judged unacceptable by flu-
4708 ent speakers of a language.) These examples cover only a tiny corner of English deriva-
4709 tional morphology, but a number of things stand out. The suffix *-ful* converts the nouns
4710 *grace* and *disgrace* into adjectives, and the suffix *-ly* converts adjectives into adverbs. These
4711 suffixes must be applied in the correct order, as shown by the unacceptability of **grace-*
4712 *lyful*. The *-ful* suffix works for only some words, as shown by the use of *alluring* as the
4713 adjectival form of *allure*. Other changes are made with prefixes, such as the derivation
4714 of *disgrace* from *grace*, which roughly corresponds to a negation; however, *fair* is negated
4715 with the *un-* prefix instead. Finally, while the first three examples suggest that the direc-
4716 tion of derivation is noun → adjective → adverb, the example of *fair* suggests that the
4717 adjective can also be the base form, with the *-ness* suffix performing the conversion to a
4718 noun.

4719 Can we build a computer program that accepts only well-formed English words, and
4720 rejects all others? This might at first seem trivial to solve with a brute-force attack: simply
4721 make a dictionary of all valid English words. But such an approach fails to account for
4722 morphological **productivity** — the applicability of existing morphological rules to new
4723 words and names, such as *Trump* to *Trumpy* and *Trumpkin*, and *Clinton* to *Clintonian* and
4724 *Clintonite*. We need an approach that represents morphological rules explicitly, and for
4725 this we will try a finite state acceptor.

4726 The dictionary approach can be implemented as a finite state acceptor, with the vo-
4727 cabulary Σ equal to the vocabulary of English, and a transition from the start state to the
4728 accepting state for each word. But this would of course fail to generalize beyond the origi-
4729 nal vocabulary, and would not capture anything about the **morphotactic** rules that govern
4730 derivations from new words. The first step towards a more general approach is shown in
4731 Figure 9.2, which is the state diagram for a finite state acceptor in which the vocabulary
4732 consists of **morphemes**, which include **stems** (e.g., *grace*, *allure*) and **affixes** (e.g., *dis-*, *-ing*,
4733 *-ly*). This finite state acceptor consists of a set of paths leading away from the start state,
4734 with derivational affixes added along the path. Except for q_{neg} , the states on these paths
4735 are all final, so the FSA will accept *disgrace*, *disgraceful*, and *disgracefully*, but not *dis-*.

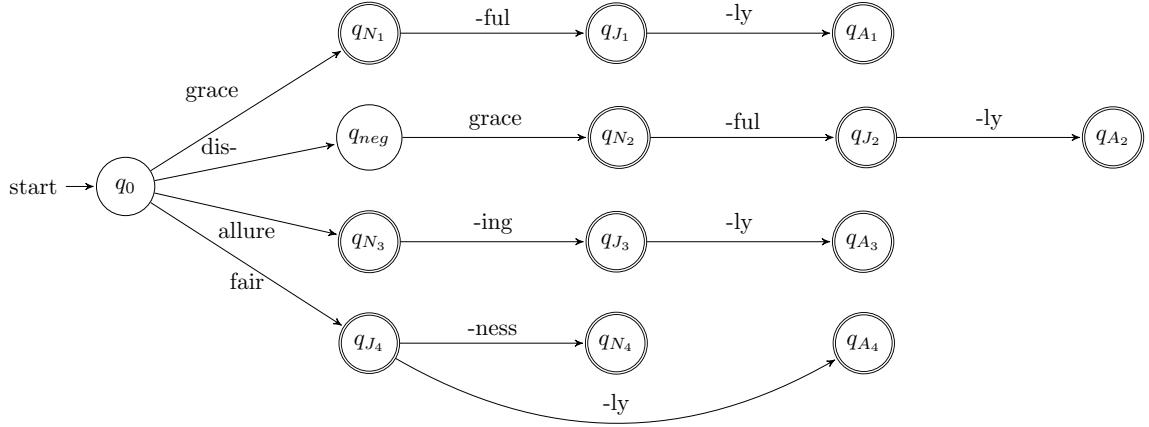


Figure 9.2: A finite state acceptor for a fragment of English derivational morphology. Each path represents possible derivations from a single root form.

4736 This FSA can be **minimized** to the form shown in Figure 9.3, which makes the general-
 4737 ity of the finite state approach more apparent. For example, the transition from q_0 to
 4738 q_{J_2} can be made to accept not only *fair* but any single-morpheme (**monomorphemic**) ad-
 4739 jective that takes *-ness* and *-ly* as suffixes. In this way, the finite state acceptor can easily
 4740 be extended: as new word stems are added to the vocabulary, their derived forms will be
 4741 accepted automatically. Of course, this FSA would still need to be extended considerably
 4742 to cover even this small fragment of English morphology. As shown by cases like *music*
 4743 \rightarrow *musical*, *athlete* \rightarrow *athletic*, English includes several classes of nouns, each with its own
 4744 rules for derivation.

4745 The FSAs shown in Figure 9.2 and 9.3 accept *allureing*, not *alluring*. This reflects a dis-
 4746 tinction between morphology — the question of which morphemes to use, and in what
 4747 order — and **orthography** — the question of how the morphemes are rendered in written
 4748 language. Just as orthography requires dropping the *e* preceding the *-ing* suffix, **phonol-**
 4749 **ogy** imposes a related set of constraints on how words are rendered in speech. As we will
 4750 see soon, these issues can be handled by **finite state transducers**, which are finite state
 4751 automata that take inputs and produce outputs.

4752 9.1.3 Weighted finite state acceptors

4753 According to the FSA treatment of morphology, every word is either in or out of the lan-
 4754 guage, with no wiggle room. Perhaps you agree that *musicky* and *fishful* are not valid
 4755 English words; but if forced to choose, you probably find *a fishful stew* or *a musicky trib-*
 4756 *ute* preferable to *behaving disgracelyful*. Rather than asking whether a word is acceptable,
 4757 we might like to ask how acceptable it is. Aronoff (1976, page 36) puts it another way:

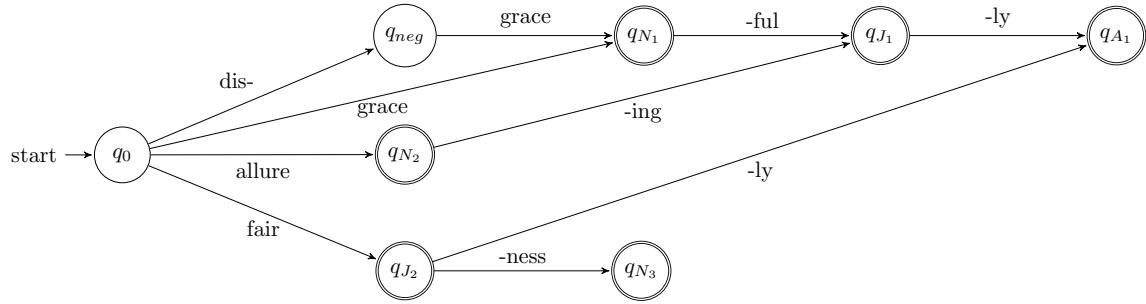


Figure 9.3: Minimization of the finite state acceptor shown in Figure 9.2.

4758 “Though many things are possible in morphology, some are more possible than others.”
 4759 But finite state acceptors give no way to express preferences among technically valid
 4760 choices.

4761 **Weighted finite state acceptors (WFSAs)** are generalizations of FSAs, in which each
 4762 accepting path is assigned a score, computed from the transitions, the initial state, and the
 4763 final state. Formally, a weighted finite state acceptor $M = (Q, \Sigma, \lambda, \rho, \delta)$ consists of:

- 4764 • a finite set of states $Q = \{q_0, q_1, \dots, q_n\}$;
- 4765 • a finite alphabet Σ of input symbols;
- 4766 • an initial weight function, $\lambda : Q \rightarrow \mathbb{R}$;
- 4767 • a final weight function $\rho : Q \rightarrow \mathbb{R}$;
- 4768 • a transition function $\delta : Q \times \Sigma \times Q \rightarrow \mathbb{R}$.

4769 WFSAs depart from the FSA formalism in three ways: every state can be an initial
 4770 state, with score $\lambda(q)$; every state can be an accepting state, with score $\rho(q)$; transitions are
 4771 possible between any pair of states on any input, with a score $\delta(q_i, \omega, q_j)$. Nonetheless,
 4772 FSAs can be viewed as a special case: for any FSA M we can build an equivalent WFSA
 4773 by setting $\lambda(q) = \infty$ for all $q \neq q_0$, $\rho(q) = \infty$ for all $q \notin F$, and $\delta(q_i, \omega, q_j) = \infty$ for all
 4774 transitions $\{(q_1, \omega) \rightarrow q_2\}$ that are not permitted by the transition function of M .

4775 The total score for any path $\pi = t_1, t_2, \dots, t_N$ is equal to the sum of these scores,

$$d(\pi) = \lambda(\text{from-state}(t_1)) + \sum_n^N \delta(t_n) + \rho(\text{to-state}(t_N)). \quad [9.5]$$

4776 A **shortest-path algorithm** is used to find the minimum-cost path through a WFSA for
 4777 string ω , with time complexity $\mathcal{O}(E + V \log V)$, where E is the number of edges and V is
 4778 the number of vertices (Cormen et al., 2009).²

²Shortest-path algorithms find the path with the minimum cost. In many cases, the path weights are log

4779 **N-gram language models as WFSAs**

4780 In **n-gram language models** (see § 6.1), the probability of a sequence of tokens w_1, w_2, \dots, w_M
 4781 is modeled as,

$$p(w_1, \dots, w_M) \approx \prod_{m=1}^M p_n(w_m | w_{m-1}, \dots, w_{m-n+1}). \quad [9.6]$$

The log probability under an n -gram language model can be modeled in a WFSA. First consider a unigram language model. We need only a single state q_0 , with transition scores $\delta(q_0, \omega, q_0) = \log p_1(\omega)$. The initial and final scores can be set to zero. Then the path score for w_1, w_2, \dots, w_M is equal to,

$$0 + \sum_m^M \delta(q_0, w_m, q_0) + 0 = \sum_m^M \log p_1(w_m). \quad [9.7]$$

For an n -gram language model with $n > 1$, we need probabilities that condition on the past history. For example, in a bigram language model, the transition weights must represent $\log p_2(w_m | w_{m-1})$. The transition scoring function must somehow “remember” the previous word or words. This can be done by adding more states: to model the bigram probability $p_2(w_m | w_{m-1})$, we need a state for every possible w_{m-1} — a total of V states. The construction indexes each state q_i by a context event $w_{m-1} = i$. The weights are then assigned as follows:

$$\begin{aligned} \delta(q_i, \omega, q_j) &= \begin{cases} \log \Pr(w_m = j | w_{m-1} = i), & \omega = j \\ -\infty, & \omega \neq j \end{cases} \\ \lambda(q_i) &= \log \Pr(w_1 = i | w_0 = \square) \\ \rho(q_i) &= \log \Pr(w_{M+1} = \blacksquare | w_M = i). \end{aligned}$$

4782 The transition function is designed to ensure that the context is recorded accurately:
 4783 we can move to state j on input ω only if $\omega = j$; otherwise, transitioning to state j is
 4784 forbidden by the weight of $-\infty$. The initial weight function $\lambda(q_i)$ is the log probability of
 4785 receiving i as the first token, and the final weight function $\rho(q_i)$ is the log probability of
 4786 receiving an “end-of-string” token after observing $w_M = i$.

4787 ***Semiring weighted finite state acceptors**

4788 The n -gram language model WFSA is deterministic: each input has exactly one accepting
 4789 path, for which the WFSA computes a score. In non-deterministic WFSAs, a given input

probabilities, so we want the path with the maximum score, which can be accomplished by making each local score into a *negative* log-probability.

4790 may have multiple accepting paths. In some applications, the score for the input is ag-
 4791 gregated across all such paths. Such aggregate scores can be computed by generalizing
 4792 WFSAs with **semiring notation**, first introduced in § 7.7.3.

4793 Let $d(\pi)$ represent the total score for path $\pi = t_1, t_2, \dots, t_N$, which is computed as,

$$d(\pi) = \lambda(\text{from-state}(t_1)) \otimes \delta(t_1) \otimes \delta(t_2) \otimes \dots \otimes \delta(t_N) \otimes \rho(\text{to-state}(t_N)). \quad [9.8]$$

4794 This is a generalization of Equation 9.5 to semiring notation, using the semiring multipli-
 4795 cation operator \otimes in place of addition.

4796 Now let $s(\omega)$ represent the total score for all paths $\Pi(\omega)$ that consume input ω ,

$$s(\omega) = \bigoplus_{\pi \in \Pi(\omega)} d(\pi). \quad [9.9]$$

4797 Here, semiring addition (\oplus) is used to combine the scores of multiple paths.

4798 The generalization to semirings covers a number of useful special cases. In the log-
 4799 probability semiring, multiplication is defined as $\log p(x) \otimes \log p(y) = \log p(x) + \log p(y)$,
 4800 and addition is defined as $\log p(x) \oplus \log p(y) = \log(p(x) + p(y))$. Thus, $s(\omega)$ represents
 4801 the log-probability of accepting input ω , marginalizing over all paths $\pi \in \Pi(\omega)$. In the
 4802 **boolean semiring**, the \otimes operator is logical conjunction, and the \oplus operator is logical
 4803 disjunction. This reduces to the special case of unweighted finite state acceptors, where
 4804 the score $s(\omega)$ is a boolean indicating whether there exists any accepting path for ω . In
 4805 the **tropical semiring**, the \oplus operator is a maximum, so the resulting score is the score of
 4806 the best-scoring path through the WFSAs. The OPENFST toolkit uses semirings and poly-
 4807 morphism to implement general algorithms for weighted finite state automata (Allauzen
 4808 et al., 2007).

4809 *Interpolated n -gram language models

4810 Recall from § 6.2.3 that an **interpolated n -gram language model** combines the probabili-
 4811 ties from multiple n -gram models. For example, an interpolated bigram language model
 4812 computes the probability,

$$\hat{p}(w_m | w_{m-1}) = \lambda_1 p_1(w_m) + \lambda_2 p_2(w_m | w_{m-1}), \quad [9.10]$$

4813 with \hat{p} indicating the interpolated probability, p_2 indicating the bigram probability, and
 4814 p_1 indicating the unigram probability. Setting $\lambda_2 = (1 - \lambda_1)$ ensures that the probabilities
 4815 sum to one.

4816 Interpolated bigram language models can be implemented using a non-deterministic
 4817 WFSAs (Knight and May, 2009). The basic idea is shown in Figure 9.4. In an interpolated
 4818 bigram language model, there is one state for each element in the vocabulary — in this

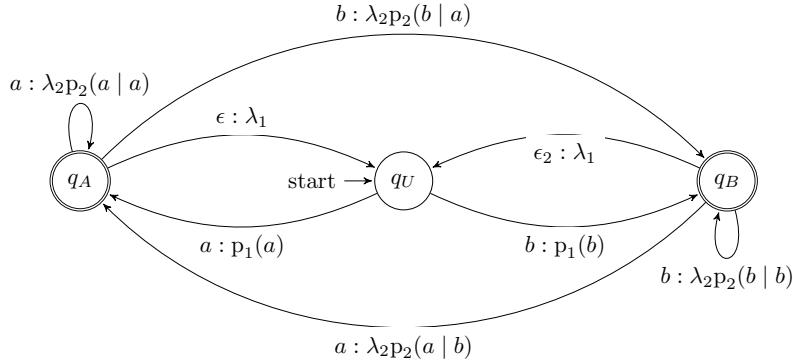


Figure 9.4: WFSA implementing an interpolated bigram/unigram language model, on the alphabet $\Sigma = \{a, b\}$. For simplicity, the WFSA is constrained to force the first token to be generated from the unigram model, and does not model the emission of the end-of-sequence token.

4819 case, the states q_A and q_B — which capture the contextual conditioning in the bigram
 4820 probabilities. To model unigram probabilities, there is an additional state q_U , which “forgets”
 4821 the context. Transitions out of q_U involve unigram probabilities, $p_1(a)$ and $p_2(b)$;
 4822 transitions into q_U emit the empty symbol ϵ , and have probability λ_1 , reflecting the inter-
 4823 polation weight for the unigram model. The interpolation weight for the bigram model is
 4824 included in the weight of the transition $q_A \rightarrow q_B$.

4825 The epsilon transitions into q_U make this WFSA non-deterministic. Consider the score
 4826 for the sequence (a, b, b) . The initial state is q_U , so the symbol a is generated with score
 4827 $p_1(a)$ ³ Next, we can generate b from the unigram model by taking the transition $q_A \rightarrow q_B$,
 4828 with score $\lambda_2 p_2(b | a)$. Alternatively, we can take a transition back to q_U with score λ_1 ,
 4829 and then emit b from the unigram model with score $p_1(b)$. To generate the final b token,
 4830 we face the same choice: emit it directly from the self-transition to q_B , or transition to q_U
 4831 first.

The total score for the sequence (a, b, b) is the semiring sum over all accepting paths,

$$\begin{aligned}
 s(a, b, b) &= (p_1(a) \otimes \lambda_2 p_2(b | a) \otimes \lambda_2 p_2(b | b)) \\
 &\oplus (p_1(a) \otimes \lambda_1 \otimes p_1(b) \otimes \lambda_2 p_2(b | b)) \\
 &\oplus (p_1(a) \otimes \lambda_2 p_2(b | a) \otimes p_1(b) \otimes p_1(b)) \\
 &\oplus (p_1(a) \otimes \lambda_1 \otimes p_1(b) \otimes p_1(b) \otimes p_1(b)). \tag{[9.11]}
 \end{aligned}$$

4832 Each line in Equation 9.11 represents the probability of a specific path through the WFSA.
 4833 In the probability semiring, \otimes is multiplication, so that each path is the product of each

³We could model the sequence-initial bigram probability $p_2(a | \square)$, but for simplicity the WFSA does not admit this possibility, which would require another state.

4834 transition weight, which are themselves probabilities. The \oplus operator is addition, so that
 4835 the total score is the sum of the scores (probabilities) for each path. This corresponds to
 4836 the probability under the interpolated bigram language model.

4837 **9.1.4 Finite state transducers**

4838 Finite state acceptors can determine whether a string is in a regular language, and weighted
 4839 finite state acceptors can compute a score for every string over a given alphabet. **Finite**
 4840 **state transducers** (FSTs) extend the formalism further, by adding an output symbol to each
 4841 transition. Formally, a finite state transducer is a tuple $T = (Q, \Sigma, \Omega, \lambda, \rho, \delta)$, with Ω repre-
 4842 senting an output vocabulary and the transition function $\delta : Q \times (\Sigma \cup \epsilon) \times (\Omega \cup \epsilon) \times Q \rightarrow \mathbb{R}$
 4843 mapping from states, input symbols, and output symbols to states. The remaining ele-
 4844 ments (Q, Σ, λ, ρ) are identical to their definition in weighted finite state acceptors (§ 9.1.3).
 4845 Thus, each path through the FST T transduces the input string into an output.

4846 **String edit distance**

The **edit distance** between two strings s and t is a measure of how many operations are required to transform one string into another. There are several ways to compute edit distance, but one of the most popular is the Levenshtein edit distance, which counts the minimum number of insertions, deletions, and substitutions. This can be computed by a one-state weighted finite state transducer, in which the input and output alphabets are identical. For simplicity, consider the alphabet $\Sigma = \Omega = \{a, b\}$. The edit distance can be computed by a one-state transducer with the following transitions,

$$\delta(q, a, a, q) = \delta(q, b, b, q) = 0 \quad [9.12]$$

$$\delta(q, a, b, q) = \delta(q, b, a, q) = 1 \quad [9.13]$$

$$\delta(q, a, \epsilon, q) = \delta(q, b, \epsilon, q) = 1 \quad [9.14]$$

$$\delta(q, \epsilon, a, q) = \delta(q, \epsilon, b, q) = 1. \quad [9.15]$$

4847 The state diagram is shown in Figure 9.5.

4848 For a given string pair, there are multiple paths through the transducer: the best-
 4849 scoring path from *dessert* to *desert* involves a single deletion, for a total score of 1; the
 4850 worst-scoring path involves seven deletions and six additions, for a score of 13.

4851 **The Porter stemmer**

The Porter (1980) stemming algorithm is a “lexicon-free” algorithm for stripping suffixes from English words, using a sequence of character-level rules. Each rule can be described

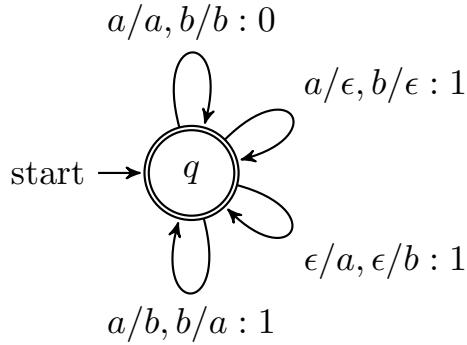


Figure 9.5: State diagram for the Levenshtein edit distance finite state transducer. The label $x/y : c$ indicates a cost of c for a transition with input x and output y .

by an unweighted finite state transducer. The first rule is:

-sses → -ss e.g., *dresses* → *dress* [9.16]

-ies → -i e.g., *parties* → *parti* [9.17]

-ss → -ss e.g., *dress* → *dress* [9.18]

-s → ε e.g., *cats* → *cat* [9.19]

4852 The final two lines appear to conflict; they are meant to be interpreted as an instruction
 4853 to remove a terminal *-s* unless it is part of an *-ss* ending. A state diagram to handle just
 4854 these final two lines is shown in Figure 9.6. Make sure you understand how this finite
 4855 state transducer handles *cats*, *steps*, *bass*, and *basses*.

4856 Inflectional morphology

4857 In **inflectional morphology**, word **lemmas** are modified to add grammatical information
 4858 such as tense, number, and case. For example, many English nouns are pluralized by the
 4859 suffix *-s*, and many verbs are converted to past tense by the suffix *-ed*. English's inflectional
 4860 morphology is considerably simpler than many of the world's languages. For example,
 4861 Romance languages (derived from Latin) feature complex systems of verb suffixes which
 4862 must agree with the person and number of the verb, as shown in Table 9.1.

4863 The task of morphological analysis is to read a form like *canto*, and output an analysis
 4864 like CANTAR+VERB+PRESIND+1P+SING, where +PRESIND describes the tense as present
 4865 indicative, +1P indicates the first-person, and +SING indicates the singular number. The
 4866 task of morphological generation is the reverse, going from CANTAR+VERB+PRESIND+1P+SING
 4867 to *canto*. Finite state transducers are an attractive solution, because they can solve both
 4868 problems with a single model (Beesley and Karttunen, 2003). As an example, Figure 9.7
 4869 shows a fragment of a finite state transducer for Spanish inflectional morphology. The

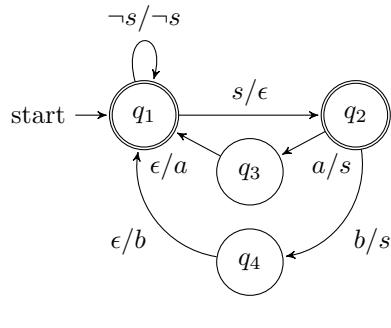


Figure 9.6: State diagram for final two lines of step 1a of the Porter stemming diagram. States q_3 and q_4 “remember” the observations a and b respectively; the ellipsis \dots represents additional states for each symbol in the input alphabet. The notation $\neg s / \neg s$ is not part of the FST formalism; it is a shorthand to indicate a set of self-transition arcs for every input/output symbol except s .

infinitive	cantar (to sing)	comer (to eat)	vivir (to live)
yo (1st singular)	canto	como	vivo
tu (2nd singular)	cantas	comes	vives
él, ella, usted (3rd singular)	canta	come	vive
nosotros (1st plural)	cantamos	comemos	vivimos
vosotros (2nd plural, informal)	cantáis	coméis	vívís
ellos, ellas (3rd plural); ustedes (2nd plural)	cantan	comen	viven

Table 9.1: Spanish verb inflections for the present indicative tense. Each row represents a person and number, and each column is a regular example from a class of verbs, as indicated by the ending of the infinitive form.

4870 input vocabulary Σ corresponds to the set of letters used in Spanish spelling, and the out-
 4871 put vocabulary Ω corresponds to these same letters, plus the vocabulary of morphological
 4872 features (e.g., +SING, +VERB). In Figure 9.7, there are two paths that take *canto* as input,
 4873 corresponding to the verb and noun meanings; the choice between these paths could be
 4874 guided by a part-of-speech tagger. By **inversion**, the inputs and outputs for each trans-
 4875ition are switched, resulting in a finite state generator, capable of producing the correct
 4876 **surface form** for any morphological analysis.

4877 Finite state morphological analyzers and other unweighted transducers can be de-
 4878 signed by hand. The designer’s goal is to avoid **overgeneration** — accepting strings or
 4879 making transductions that are not valid in the language — as well as **undergeneration**

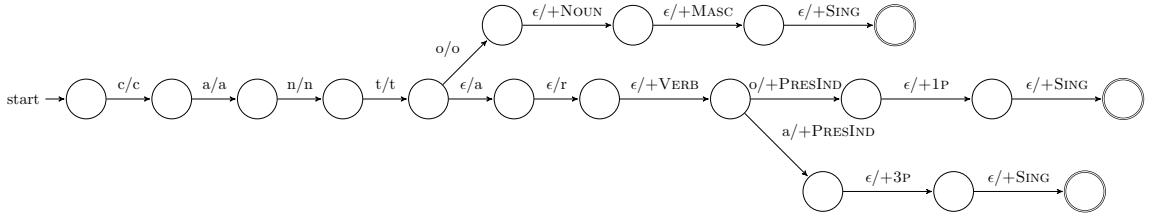


Figure 9.7: Fragment of a finite state transducer for Spanish morphology. There are two accepting paths for the input *canto*: *canto*+NOUN+MASC+SING (masculine singular noun, meaning a song), and *cantar*+VERB+PRESIND+1P+SING (I sing). There is also an accepting path for *canta*, with output *cantar*+VERB+PRESIND+3P+SING (he/she sings).

4880 — failing to accept strings or transductions that are valid. For example, a pluralization
 4881 transducer that does not accept *foot/feet* would undergenerate. Suppose we “fix” the trans-
 4882 ducer to accept this example, but as a side effect, it now accepts *boot/beet*; the transducer
 4883 would then be said to overgenerate. If a transducer accepts *foot/foots* but not *foot/feet*, then
 4884 it simultaneously overgenerates and undergenerates.

4885 Finite state composition

4886 Designing finite state transducers to capture the full range of morphological phenomena
 4887 in any real language is a huge task. Modularization is a classic computer science approach
 4888 for this situation: decompose a large and unwieldy problem into a set of subproblems,
 4889 each of which will hopefully have a concise solution. Finite state automata can be mod-
 4890 ularized through **composition**: feeding the output of one transducer T_1 as the input to
 4891 another transducer T_2 , written $T_2 \circ T_1$. Formally, if there exists some y such that $(x, y) \in T_1$
 4892 (meaning that T_1 produces output y on input x), and $(y, z) \in T_2$, then $(x, z) \in (T_2 \circ T_1)$.
 4893 Because finite state transducers are closed under composition, there is guaranteed to be
 4894 a single finite state transducer that $T_3 = T_2 \circ T_1$, which can be constructed as a machine
 4895 with one state for each pair of states in T_1 and T_2 (Mohri et al., 2002).

4896 **Example: Morphology and orthography** In English morphology, the suffix *-ed* is added
 4897 to signal the past tense for many verbs: *cook*→*cooked*, *want*→*wanted*, etc. However, English
 4898 **orthography** dictates that this process cannot produce a spelling with consecutive e’s, so
 4899 that *bake*→*baked*, not *bakeed*. A modular solution is to build separate transducers for mor-
 4900 phology and orthography. The morphological transducer T_M transduces from *bake*+PAST
 4901 to *bake+ed*, with the + symbol indicating a segment boundary. The input alphabet of T_M
 4902 includes the lexicon of words and the set of morphological features; the output alphabet
 4903 includes the characters *a-z* and the + boundary marker. Next, an orthographic transducer
 4904 T_O is responsible for the transductions *cook+ed*→*cooked*, and *bake+ed*→*baked*. The input
 4905 alphabet of T_O must be the same as the output alphabet for T_M , and the output alphabet

4906 is simply the characters *a-z*. The composed transducer ($T_O \circ T_M$) then transduces from
 4907 *bake*+PAST to the spelling *baked*. The design of T_O is left as an exercise.

Example: Hidden Markov models Hidden Markov models (chapter 7) can be viewed as weighted finite state transducers, and they can be constructed by transduction. Recall that a hidden Markov model defines a joint probability over words and tags, $p(w, y)$, which can be computed as a path through a **trellis** structure. This trellis is itself a weighted finite state acceptor, with edges between all adjacent nodes $q_{m-1,i} \rightarrow q_{m,j}$ on input $Y_m = j$. The edge weights are log-probabilities,

$$\delta(q_{m-1,i}, Y_m = j, q_{m,j}) = \log p(w_m, Y_m = j \mid Y_{m-1} = i) \quad [9.20]$$

$$= \log p(w_m \mid Y_m = j) + \log \Pr(Y_m = j \mid Y_{m-1} = i). \quad [9.21]$$

4908 Because there is only one possible transition for each tag Y_m , this WFSA is deterministic.
 4909 The score for any tag sequence $\{y_m\}_{m=1}^M$ is the sum of these log-probabilities, correspond-
 4910 ing to the total log probability $\log p(w, y)$. Furthermore, the trellis can be constructed by
 4911 the composition of simpler FSTs.

- 4912 • First, construct a “transition” transducer to represent a bigram probability model
 4913 over tag sequences, T_T . This transducer is almost identical to the n -gram language
 4914 model acceptor in § 9.1.3: there is one state for each tag, and the edge weights equal
 4915 to the transition log-probabilities, $\delta(q_i, j, j, q_j) = \log \Pr(Y_m = j \mid Y_{m-1} = i)$. Note
 4916 that T_T is a transducer, with identical input and output at each arc; this makes it
 4917 possible to compose T_T with other transducers.
- 4918 • Next, construct an “emission” transducer to represent the probability of words given
 4919 tags, T_E . This transducer has only a single state, with arcs for each word/tag pair,
 4920 $\delta(q_0, i, j, q_0) = \log \Pr(W_m = j \mid Y_m = i)$. The input vocabulary is the set of all tags,
 4921 and the output vocabulary is the set of all words.
- 4922 • The composition $T_E \circ T_T$ is a finite state transducer with one state per tag, as shown
 4923 in Figure 9.8. Each state has $V \times K$ outgoing edges, representing transitions to each
 4924 of the K other states, with outputs for each of the V words in the vocabulary. The
 4925 weights for these edges are equal to,

$$\delta(q_i, Y_m = j, w_m, q_j) = \log p(w_m, Y_m = j \mid Y_{m-1} = i). \quad [9.22]$$

- 4926 • The trellis is a structure with $M \times K$ nodes, for each of the M words to be tagged and
 4927 each of the K tags in the tagset. It can be built by composition of $(T_E \circ T_T)$ against an
 4928 unweighted **chain FSA** $M_A(w)$ that is specially constructed to accept only a given
 4929 input w_1, w_2, \dots, w_M , shown in Figure 9.9. The trellis for input w is built from the
 4930 composition $M_A(w) \circ (T_E \circ T_T)$. Composing with the unweighted $M_A(w)$ does not
 4931 affect the edge weights from $(T_E \circ T_T)$, but it selects the subset of paths that generate
 4932 the word sequence w .

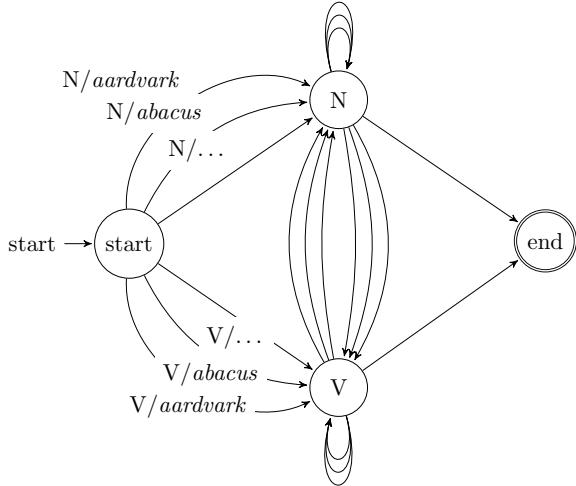


Figure 9.8: Finite state transducer for hidden Markov models, with a small tagset of nouns and verbs. For each pair of tags (including self-loops), there is an edge for every word in the vocabulary. For simplicity, input and output are only shown for the edges from the start state. Weights are also omitted from the diagram; for each edge from q_i to q_j , the weight is equal to $\log p(w_m, Y_m = j \mid Y_{m-1} = i)$, except for edges to the end state, which are equal to $\log \Pr(Y_m = \diamond \mid Y_{m-1} = i)$.

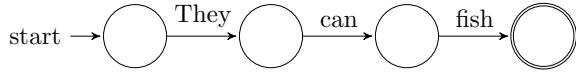


Figure 9.9: Chain finite state acceptor for the input *They can fish*.

4933 9.1.5 *Learning weighted finite state automata

4934 In generative models such as n -gram language models and hidden Markov models, the
 4935 edge weights correspond to log probabilities, which can be obtained from relative fre-
 4936 quency estimation. However, in other cases, we wish to learn the edge weights from in-
 4937 put/output pairs. This is difficult in non-deterministic finite state automata, because we
 4938 do not observe the specific arcs that are traversed in accepting the input, or in transducing
 4939 from input to output. The path through the automaton is a **latent variable**.

4940 Chapter 5 presented one method for learning with latent variables: expectation max-
 4941 imization (EM). This involves computing a distribution $q(\cdot)$ over the latent variable, and
 4942 iterating between updates to this distribution and updates to the parameters — in this
 4943 case, the arc weights. The **forward-backward algorithm** (§ 7.5.3) describes a dynamic
 4944 program for computing a distribution over arcs in the trellis structure of a hidden Markov

model, but this is a special case of the more general problem for finite state automata. Eisner (2002) describes an **expectation semiring**, which enables the expected number of transitions across each arc to be computed through a semiring shortest-path algorithm. Alternative approaches for generative models include Markov Chain Monte Carlo (Chiang et al., 2010) and spectral learning (Balle et al., 2011).

Further afield, we can take a perceptron-style approach, with each arc corresponding to a feature. The classic perceptron update would update the weights by subtracting the difference between the feature vector corresponding to the predicted path and the feature vector corresponding to the correct path. Since the path is not observed, we resort to a **latent variable perceptron**. The model is described formally in § 12.4, but the basic idea is to compute an update from the difference between the features from the predicted path and the features for the best-scoring path that generates the correct output.

9.2 Context-free languages

Beyond the class of regular languages lie the context-free languages. An example of a language that is context-free but not finite state is the set of arithmetic expressions with balanced parentheses. Intuitively, to accept only strings in this language, an FSA would have to “count” the number of left parentheses, and make sure that they are balanced against the number of right parentheses. An arithmetic expression can be arbitrarily long, yet by definition an FSA has a finite number of states. Thus, for any FSA, there will be a string that with too many parentheses to count. More formally, the **pumping lemma** is a proof technique for showing that languages are not regular. It is typically demonstrated for the simpler case $a^n b^n$, the language of strings containing a sequence of a 's, and then an equal-length sequence of b 's.⁴

There are at least two arguments for the relevance of non-regular formal languages to linguistics. First, there are natural language phenomena that are argued to be isomorphic to $a^n b^n$. For English, the classic example is **center embedding**, shown in Figure 9.10. The initial expression *the dog* specifies a single dog. Embedding this expression into *the cat ... chased* specifies a particular cat — the one chased by the dog. This cat can then be embedded again to specify a goat, in the less felicitous but arguably grammatical expression, *the goat the cat the dog chased kissed*, which refers to the goat who was kissed by the cat which was chased by the dog. Chomsky (1957) argues that to be grammatical, a center-embedded construction must be balanced: if it contains n noun phrases (e.g., *the cat*), they must be followed by exactly $n - 1$ verbs. An FSA that could recognize such expressions would also be capable of recognizing the language $a^n b^n$. Because we can prove that no FSA exists for $a^n b^n$, no FSA can exist for center embedded constructions either. En-

⁴Details of the proof can be found in an introductory computer science theory textbook (e.g., Sipser, 2012).

			the dog	
	the cat	the dog	chased	
the goat	the cat	the dog	chased	kissed
			...	

Figure 9.10: Three levels of center embedding

4980 glish includes center embedding, and so the argument goes, English grammar as a whole
 4981 cannot be regular.⁵

4982 A more practical argument for moving beyond regular languages is modularity. Many
 4983 linguistic phenomena — especially in syntax — involve constraints that apply at long
 4984 distance. Consider the problem of determiner-noun number agreement in English: we
 4985 can say *the coffee* and *these coffees*, but not **these coffee*. By itself, this is easy enough to model
 4986 in an FSA. However, fairly complex modifying expressions can be inserted between the
 4987 determiner and the noun:

- 4988 (9.2) a. the burnt coffee
 4989 b. the badly-ground coffee
 4990 c. the burnt and badly-ground Italian coffee
 4991 d. these burnt and badly-ground Italian coffees
 4992 e. * these burnt and badly-ground Italian coffee

4993 Again, an FSA can be designed to accept modifying expressions such as *burnt and badly-*
 4994 *ground Italian*. Let's call this FSA F_M . To reject the final example, a finite state acceptor
 4995 must somehow "remember" that the determiner was plural when it reaches the noun *cof-*
 4996 *fee* at the end of the expression. The only way to do this is to make two identical copies
 4997 of F_M : one for singular determiners, and one for plurals. While this is possible in the
 4998 finite state framework, it is inconvenient — especially in languages where more than one
 4999 attribute of the noun is marked by the determiner. **Context-free languages** facilitate mod-
 5000 ularity across such long-range dependencies.

5001 9.2.1 Context-free grammars

5002 Context-free languages are specified by **context-free grammars** (CFGs), which are tuples
 5003 (N, Σ, R, S) consisting of:

⁵The claim that arbitrarily deep center-embedded expressions are grammatical has drawn skepticism. Corpus evidence shows that embeddings of depth greater than two are exceedingly rare (Karlsson, 2007), and that embeddings of depth greater than three are completely unattested. If center-embedding is capped at some finite depth, then it is regular.

$$\begin{aligned}
 S &\rightarrow S \text{ OP } S \mid \text{NUM} \\
 \text{OP} &\rightarrow + \mid - \mid \times \mid \div \\
 \text{NUM} &\rightarrow \text{NUM DIGIT} \mid \text{DIGIT} \\
 \text{DIGIT} &\rightarrow 0 \mid 1 \mid 2 \mid \dots \mid 9
 \end{aligned}$$

Figure 9.11: A context-free grammar for arithmetic expressions

- 5004 • a finite set of **non-terminals** N ;
 5005 • a finite alphabet Σ of **terminal symbols**;
 5006 • a set of **production rules** R , each of the form $A \rightarrow \beta$, where $A \in N$ and $\beta \in (\Sigma \cup N)^*$;
 5007 • a designated start symbol S .

5008 In the production rule $A \rightarrow \beta$, the left-hand side (LHS) A must be a non-terminal;
 5009 the right-hand side (RHS) can be a sequence of terminals or non-terminals, $\{n, \sigma\}^*, n \in$
 5010 $N, \sigma \in \Sigma$. A non-terminal can appear on the left-hand side of many production rules.
 5011 A non-terminal can appear on both the left-hand side and the right-hand side; this is a
 5012 **recursive production**, and is analogous to self-loops in finite state automata. The name
 5013 “context-free” is based on the property that the production rule depends only on the LHS,
 5014 and not on its ancestors or neighbors; this is analogous to Markov property of finite state
 5015 automata, in which the behavior at each step depends only on the current state, on not on
 5016 the path by which that state was reached.

5017 A **derivation** τ is a sequence of steps from the start symbol S to a surface string $w \in \Sigma^*$,
 5018 which is the **yield** of the derivation. A string w is in a context-free language if there is
 5019 some derivation from S yielding w . **Parsing** is the problem of finding a derivation for a
 5020 string in a grammar. Algorithms for parsing are described in chapter 10.

5021 Like regular expressions, context-free grammars define the language but not the com-
 5022 putation necessary to recognize it. The context-free analogues to finite state acceptors are
 5023 **pushdown automata**, a theoretical model of computation in which input symbols can be
 5024 pushed onto a stack with potentially infinite depth. For more details, see Sipser (2012).

5025 Example

5026 Figure 9.11 shows a context-free grammar for arithmetic expressions such as $1 + 2 \div 3 - 4$.
 5027 In this grammar, the terminal symbols include the digits $\{1, 2, \dots, 9\}$ and the op-
 5028 erators $\{+, -, \times, \div\}$. The rules include the $|$ symbol, a notational convenience that makes
 5029 it possible to specify multiple right-hand sides on a single line: the statement $A \rightarrow x | y$

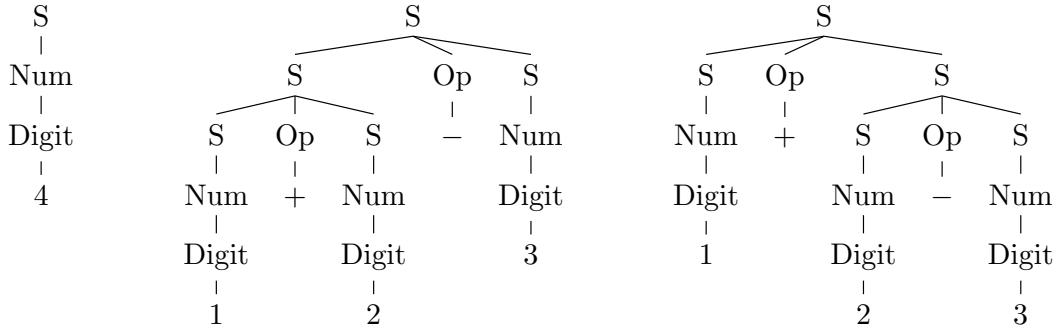


Figure 9.12: Some example derivations from the arithmetic grammar in Figure 9.11

5030 defines *two* productions, $A \rightarrow x$ and $A \rightarrow y$. This grammar is recursive: the non-termals S
5031 and NUM can produce themselves.

5032 Derivations are typically shown as trees, with production rules applied from the top
5033 to the bottom. The tree on the left in Figure 9.12 describes the derivation of a single digit,
5034 through the sequence of productions $S \rightarrow \text{NUM} \rightarrow \text{DIGIT} \rightarrow 4$ (these are all **unary produc-**
5035 **tions**, because the right-hand side contains a single element). The other two trees in
5036 Figure 9.12 show alternative derivations of the string $1 + 2 - 3$. The existence of multiple
5037 derivations for a string indicates that the grammar is **ambiguous**.

Context-free derivations can also be written out according to the pre-order tree traversal.⁶ For the two derivations of $1 + 2 - 3$ in Figure 9.12, the notation is:

$$(S (S (S (\text{Num} (Digit 1))) (\text{Op} +) (S (\text{Num} (Digit 2))))) (\text{Op} -) (S (\text{Num} (Digit 3)))) \quad [9.23]$$

$$(S (S (\text{Num} (Digit 1))) (\text{Op} +) (S (\text{Num} (Digit 2)) (\text{Op} -) (S (\text{Num} (Digit 3)))))). \quad [9.24]$$

5038 Grammar equivalence and Chomsky Normal Form

A single context-free language can be expressed by more than one context-free grammar. For example, the following two grammars both define the language $a^n b^n$ for $n > 0$.

$$\begin{aligned} S &\rightarrow aSb \mid ab \\ S &\rightarrow aSb \mid aabb \mid ab \end{aligned}$$

5039 Two grammars are **weakly equivalent** if they generate the same strings. Two grammars
5040 are **strongly equivalent** if they generate the same strings via the same derivations. The
5041 grammars above are only weakly equivalent.

⁶This is a depth-first left-to-right search that prints each node the first time it is encountered (Cormen et al., 2009, chapter 12).

In **Chomsky Normal Form (CNF)**, the right-hand side of every production includes either two non-terminals, or a single terminal symbol:

$$A \rightarrow BC$$

$$A \rightarrow a$$

- 5042 All CFGs can be converted into a CNF grammar that is weakly equivalent. To convert a
 5043 grammar into CNF, we first address productions that have more than two non-terminals
 5044 on the RHS by creating new “dummy” non-terminals. For example, if we have the pro-
 5045 duction,

$$W \rightarrow X Y Z, \quad [9.25]$$

it is replaced with two productions,

$$W \rightarrow X W \setminus X \quad [9.26]$$

$$W \setminus X \rightarrow Y Z. \quad [9.27]$$

- 5046 In these productions, $W \setminus X$ is a new dummy non-terminal. This transformation **binarizes**
 5047 the grammar, which is critical for efficient bottom-up parsing, as we will see in chapter 10.
 5048 Productions whose right-hand side contains a mix of terminal and non-terminal symbols
 5049 can be replaced in a similar fashion.

- 5050 Unary non-terminal productions $A \rightarrow B$ are replaced as follows: for each production
 5051 $B \rightarrow \alpha$ in the grammar, add a new production $A \rightarrow \alpha$. For example, in the grammar
 5052 described in Figure 9.11, we would replace $\text{NUM} \rightarrow \text{DIGIT}$ with $\text{NUM} \rightarrow 1 \mid 2 \mid \dots \mid 9$.
 5053 However, we keep the production $\text{NUM} \rightarrow \text{NUM DIGIT}$, which is a valid binary produc-
 5054 tion.

5055 9.2.2 Natural language syntax as a context-free language

- 5056 Context-free grammars can be used to represent **syntax**, which is the set of rules that
 5057 determine whether an utterance is judged to be grammatical. If this representation were
 5058 perfectly faithful, then a natural language such as English could be transformed into a
 5059 formal language, consisting of exactly the (infinite) set of strings that would be judged to
 5060 be grammatical by a fluent English speaker. We could then build parsing software that
 5061 would automatically determine if a given utterance were grammatical.⁷

- 5062 Contemporary theories generally do *not* consider natural languages to be context-free
 5063 (see § 9.3), yet context-free grammars are widely used in natural language parsing. The
 5064 reason is that context-free representations strike a good balance: they cover a broad range
 5065 of syntactic phenomena, and they can be parsed efficiently. This section therefore de-
 5066 scribes how to handle a core fragment of English syntax in context-free form, following

⁷To move beyond this cursory treatment of syntax, consult the short introductory manuscript by Bender (2013), or the longer text by Akmajian et al. (2010).

5067 the conventions of the **Penn Treebank** (PTB; Marcus et al., 1993), a large-scale annotation
 5068 of English language syntax. The generalization to “mildly” context-sensitive languages is
 5069 discussed in § 9.3.

5070 The Penn Treebank annotation is a **phrase-structure grammar** of English. This means
 5071 that sentences are broken down into **constituents**, which are contiguous sequences of
 5072 words that function as coherent units for the purpose of linguistic analysis. Constituents
 5073 generally have a few key properties:

5074 **Movement.** Constituents can often be moved around sentences as units.

- 5075 (9.3) a. Abigail gave (her brother) (a fish).
 5076 b. Abigail gave (a fish) to (her brother).

5077 In contrast, *gave her* and *brother a* cannot easily be moved while preserving gram-
 5078 maticality.

5079 **Substitution.** Constituents can be substituted by other phrases of the same type.

- 5080 (9.4) a. Max thanked (his older sister).
 5081 b. Max thanked (her).

5082 In contrast, substitution is not possible for other contiguous units like *Max thanked*
 5083 and *thanked his*.

5084 **Coordination.** Coordinators like *and* and *or* can conjoin constituents.

- 5085 (9.5) a. (Abigail) and (her younger brother) bought a fish.
 5086 b. Abigail (bought a fish) and (gave it to Max).
 5087 c. Abigail (bought) and (greedily ate) a fish.

5088 Units like *brother bought* and *bought a* cannot easily be coordinated.

5089 These examples argue for units such as *her brother* and *bought a fish* to be treated as con-
 5090 stituents. Other sequences of words in these examples, such as *Abigail gave* and *brother a fish*,
 5091 cannot be moved, substituted, and coordinated in these ways. In phrase-structure
 5092 grammar, constituents are nested, so that *the senator from New Jersey* contains the con-
 5093 stituent *from New Jersey*, which in turn contains *New Jersey*. The sentence itself is the max-
 5094 imal constituent; each word is a minimal constituent, derived from a unary production
 5095 from a part-of-speech tag. Between part-of-speech tags and sentences are **phrases**. In
 5096 phrase-structure grammar, phrases have a type that is usually determined by their **head**
 5097 **word**: for example, a **noun phrase** corresponds to a noun and the group of words that

5098 modify it, such as *her younger brother*; a **verb phrase** includes the verb and its modifiers,
5099 such as *bought a fish* and *greedily ate it*.

5100 In context-free grammars, each phrase type is a non-terminal, and each constituent is
5101 the substring that the non-terminal yields. Grammar design involves choosing the right
5102 set of non-terminals. Fine-grained non-terminals make it possible to represent more fine-
5103 grained linguistic phenomena. For example, by distinguishing singular and plural noun
5104 phrases, it is possible to have a grammar of English that generates only sentences that
5105 obey subject-verb agreement. However, enforcing subject-verb agreement is considerably
5106 more complicated in languages like Spanish, where the verb must agree in both person
5107 and number with subject. In general, grammar designers must trade off between **over-**
5108 **generation** — a grammar that permits ungrammatical sentences — and **undergeneration**
5109 — a grammar that fails to generate grammatical sentences. Furthermore, if the grammar is
5110 to support manual annotation of syntactic structure, it must be simple enough to annotate
5111 efficiently.

5112 9.2.3 A phrase-structure grammar for English

5113 To better understand how phrase-structure grammar works, let's consider the specific
5114 case of the Penn Treebank grammar of English. The main phrase categories in the Penn
5115 Treebank (PTB) are based on the main part-of-speech classes: noun phrase (NP), verb
5116 phrase (VP), prepositional phrase (PP), adjectival phrase (ADJP), and adverbial phrase
5117 (ADVP). The top-level category is S, which conveniently stands in for both “sentence”
5118 and the “start” symbol. **Complement clauses** (e.g., *I take the good old fashioned ground that*
5119 *the whale is a fish*) are represented by the non-terminal SBAR. The terminal symbols in
5120 the grammar are individual words, which are generated from unary productions from
5121 part-of-speech tags (the PTB tagset is described in § 8.1).

5122 This section describes some of the most common productions from the major phrase-
5123 level categories, explaining how to generate individual tag sequences. The production
5124 rules are approached in a “theory-driven” manner: first the syntactic properties of each
5125 phrase type are described, and then some of the necessary production rules are listed. But
5126 it is important to keep in mind that the Penn Treebank was produced in a “data-driven”
5127 manner. After the set of non-terminals was specified, annotators were free to analyze each
5128 sentence in whatever way seemed most linguistically accurate, subject to some high-level
5129 guidelines. The grammar of the Penn Treebank is simply the set of productions that were
5130 required to analyze the several million words of the corpus. By design, the grammar
5131 overgenerates — it does not exclude ungrammatical sentences. Furthermore, while the
5132 productions shown here cover some of the most common cases, they are only a small
5133 fraction of the several thousand different types of productions in the Penn Treebank.

5134 **Sentences**

The most common production rule for sentences is,

$$S \rightarrow NP VP \quad [9.28]$$

which accounts for simple sentences like *Abigail ate the kimchi* — as we will see, the direct object *the kimchi* is part of the verb phrase. But there are more complex forms of sentences as well:

$$S \rightarrow ADVP NP VP \quad \begin{matrix} Unfortunately Abigail ate the kimchi. \\ Abigail ate the kimchi and Max had a burger. \end{matrix} \quad [9.29]$$

$$S \rightarrow S CC S \quad \begin{matrix} Abigail ate the kimchi and Max had a burger. \\ Eat the kimchi. \end{matrix} \quad [9.30]$$

$$S \rightarrow VP \quad \begin{matrix} Eat the kimchi. \end{matrix} \quad [9.31]$$

- 5135 where ADVP is an adverbial phrase (e.g., *unfortunately*, *very unfortunately*) and CC is a
 5136 coordinating conjunction (e.g., *and*, *but*).⁸

5137 **Noun phrases**

Noun phrases refer to entities, real or imaginary, physical or abstract: *Asha*, *the steamed dumpling*, *parts and labor*, *nobody*, *the whiteness of the whale*, and *the rise of revolutionary syndicalism in the early twentieth century*. Noun phrase productions include “bare” nouns, which may optionally follow determiners, as well as pronouns:

$$NP \rightarrow NN | NNS | NNP | PRP \quad [9.32]$$

$$NP \rightarrow DET NN | DET NNS | DET NNP \quad [9.33]$$

- 5138 The tags NN, NNS, and NNP refer to singular, plural, and proper nouns; PRP refers to
 5139 personal pronouns, and DET refers to determiners. The grammar also contains terminal
 5140 productions from each of these tags, e.g., $PRP \rightarrow I | you | we | \dots$.

Noun phrases may be modified by adjectival phrases (ADJP; e.g., *the small Russian dog*) and numbers (CD; e.g., *the five pastries*), each of which may optionally follow a determiner:

$$NP \rightarrow ADJP NN | ADJP NNS | DET ADJP NN | DET ADJP NNS \quad [9.34]$$

$$NP \rightarrow CD NNS | DET CD NNS | \dots \quad [9.35]$$

Some noun phrases include multiple nouns, such as *the liberation movement* and *an antelope horn*, necessitating additional productions:

$$NP \rightarrow NN NN | NN NNS | DET NN NN | \dots \quad [9.36]$$

⁸Notice that the grammar does not include the recursive production $S \rightarrow ADVP S$. It may be helpful to think about why this production would cause the grammar to overgenerate.

5141 These multiple noun constructions can be combined with adjectival phrases and cardinal
 5142 numbers, leading to a large number of additional productions.

Recursive noun phrase productions include coordination, prepositional phrase attachment, subordinate clauses, and verb phrase adjuncts:

$NP \rightarrow NP\ Cc\ NP$	<i>e.g., the red and the black</i>	[9.37]
$NP \rightarrow NP\ PP$	<i>e.g., the President of the Georgia Institute of Technology</i>	[9.38]
$NP \rightarrow NP\ SBAR$	<i>e.g., a whale which he had wounded</i>	[9.39]
$NP \rightarrow NP\ VP$	<i>e.g., a whale taken near Shetland</i>	[9.40]

5143 These recursive productions are a major source of ambiguity, because the VP and PP non-
 5144 terminals can also generate NP children. Thus, the *the President of the Georgia Institute of*
 5145 *Technology* can be derived in two ways, as can *a whale taken near Shetland in October*.

5146 But aside from these few recursive productions, the noun phrase fragment of the Penn
 5147 Treebank grammar is relatively flat, containing a large of number of productions that go
 5148 from NP directly to a sequence of parts-of-speech. If noun phrases had more internal
 5149 structure, the grammar would need fewer rules, which, as we will see, would make pars-
 5150 ing faster and machine learning easier. Vadas and Curran (2011) propose to add additional
 5151 structure in the form of a new non-terminal called a **nominal modifier** (NML), e.g.,

- 5152 (9.6) a. (NP (NN crude) (NN oil) (NNS prices)) (PTB analysis)
 5153 b. (NP (NML (NN crude) (NN oil)) (NNS prices)) (NML-style analysis).

5154 Another proposal is to treat the determiner as the head of a **determiner phrase** (DP;
 5155 Abney, 1987). There are linguistic arguments for and against determiner phrases (e.g.,
 5156 Van Eynde, 2006). From the perspective of context-free grammar, DPs enable more struc-
 5157 tured analyses of some constituents, e.g.,

- 5158 (9.7) a. (NP (DT the) (JJ white) (NN whale)) (PTB analysis)
 5159 b. (DP (DT the) (NP (JJ white) (NN whale))) (DP-style analysis).

5160 Verb phrases

Verb phrases describe actions, events, and states of being. The PTB tagset distinguishes several classes of verb inflections: base form (VB; *she likes to snack*), present-tense third-person singular (VBD; *she snacks*), present tense but not third-person singular (VBP; *they snack*), past tense (VBD; *they snacked*), present participle (VBG; *they are snacking*), and past participle (VBN; *they had snacked*).⁹ Each of these forms can constitute a verb phrase on its

⁹This tagset is specific to English: for example, VBP is a meaningful category only because English morphology distinguishes third-person singular from all person-number combinations.

own:

$$\text{VP} \rightarrow \text{VB} \mid \text{VBD} \mid \text{VBN} \mid \text{VBG} \mid \text{VBP} \quad [9.41]$$

More complex verb phrases can be formed by a number of recursive productions, including the use of coordination, modal verbs (MD; *she should snack*), and the infinitival *to* (TO):

$\text{VP} \rightarrow \text{MD VP}$	<i>She will snack</i>	[9.42]
$\text{VP} \rightarrow \text{VBD VP}$	<i>She had snacked</i>	[9.43]
$\text{VP} \rightarrow \text{VBZ VP}$	<i>She has been snacking</i>	[9.44]
$\text{VP} \rightarrow \text{VBN VP}$	<i>She has been snacking</i>	[9.45]
$\text{VP} \rightarrow \text{TO VP}$	<i>She wants to snack</i>	[9.46]
$\text{VP} \rightarrow \text{VP CC VP}$	<i>She buys and eats many snacks</i>	[9.47]

- 5161 Each of these productions uses recursion, with the VP non-terminal appearing in both the
 5162 LHS and RHS. This enables the creation of complex verb phrases, such as *She will have*
 5163 *wanted to have been snacking*.

Transitive verbs take noun phrases as direct objects, and ditransitive verbs take two direct objects:

$\text{VP} \rightarrow \text{VBZ NP}$	<i>She teaches algebra</i>	[9.48]
$\text{VP} \rightarrow \text{VBD NP NP}$	<i>She has been teaching algebra</i>	[9.49]
$\text{VP} \rightarrow \text{VBD NP NP}$	<i>She taught her brother algebra</i>	[9.50]

These productions are *not* recursive, so a unique production is required for each verb part-of-speech. They also do not distinguish transitive from intransitive verbs, so the resulting grammar overgenerates examples like **She sleeps sushi* and **She learns Boyang algebra*. Sentences can also be direct objects:

$\text{VP} \rightarrow \text{VBZ S}$	<i>Hunter wants to eat the kimchi</i>	[9.51]
$\text{VP} \rightarrow \text{VBZ SBAR}$	<i>Hunter knows that Tristan ate the kimchi</i>	[9.52]

- 5164 The first production overgenerates, licensing sentences like **Hunter sees Tristan eats the*
 5165 *kimchi*. This problem could be addressed by designing a more specific set of sentence
 5166 non-terminals, indicating whether the main verb can be conjugated.

Verbs can also be modified by prepositional phrases and adverbial phrases:

$\text{VP} \rightarrow \text{VBZ PP}$	<i>She studies at night</i>	[9.53]
$\text{VP} \rightarrow \text{VBZ ADVP}$	<i>She studies intensively</i>	[9.54]
$\text{VP} \rightarrow \text{ADVP VBG}$	<i>She is not studying</i>	[9.55]

5167 Again, because these productions are not recursive, the grammar must include productions
 5168 for every verb part-of-speech.

A special set of verbs, known as **copula**, can take **predicative adjectives** as direct objects:

$VP \rightarrow VBZ\ ADJP$	<i>She is hungry</i>	[9.56]
$VP \rightarrow VBP\ ADJP$	<i>Success seems increasingly unlikely</i>	[9.57]

5169 The PTB does not have a special non-terminal for copular verbs, so this production generates
 5170 non-grammatical examples such as **She eats tall*.

Particles (PRT as a phrase; RP as a part-of-speech) work to create phrasal verbs:

$VP \rightarrow VB\ PRT$	<i>She told them to fuck off</i>	[9.58]
$VP \rightarrow VBD\ PRT\ NP$	<i>They gave up their ill-gotten gains</i>	[9.59]

5171 As the second production shows, particle productions are required for all configurations
 5172 of verb parts-of-speech and direct objects.

5173 Other constituents

The remaining constituents require far fewer productions. **Prepositional phrases** almost always consist of a preposition and a noun phrase,

$PP \rightarrow IN\ NP$	<i>the whiteness of the whale</i>	[9.60]
$PP \rightarrow TO\ NP$	<i>What the white whale was to Ahab, has been hinted</i>	[9.61]

Similarly, complement clauses consist of a complementizer (usually a preposition, possibly null) and a sentence,

$SBAR \rightarrow IN\ S$	<i>She said that it was spicy</i>	[9.62]
$SBAR \rightarrow S$	<i>She said it was spicy</i>	[9.63]

Adverbial phrases are usually bare adverbs ($ADVP \rightarrow RB$), with a few exceptions:

$ADVP \rightarrow RB\ RBR$	<i>They went considerably further</i>	[9.64]
$ADVP \rightarrow ADVP\ PP$	<i>They went considerably further than before</i>	[9.65]

5174 The tag RBR is a comparative adverb.

Adjectival phrases extend beyond bare adjectives ($\text{ADJP} \rightarrow \text{JJ}$) in a number of ways:

$\text{ADJP} \rightarrow \text{RB JJ}$	<i>very hungry</i>	[9.66]
$\text{ADJP} \rightarrow \text{RBR JJ}$	<i>more hungry</i>	[9.67]
$\text{ADJP} \rightarrow \text{JJS JJ}$	<i>best possible</i>	[9.68]
$\text{ADJP} \rightarrow \text{RB JJR}$	<i>even bigger</i>	[9.69]
$\text{ADJP} \rightarrow \text{JJ CC JJ}$	<i>high and mighty</i>	[9.70]
$\text{ADJP} \rightarrow \text{JJ JJ}$	<i>West German</i>	[9.71]
$\text{ADJP} \rightarrow \text{RB VBN}$	<i>previously reported</i>	[9.72]

5175 The tags JJR and JJS refer to comparative and superlative adjectives respectively.

All of these phrase types can be coordinated:

$\text{PP} \rightarrow \text{PP CC PP}$	<i>on time and under budget</i>	[9.73]
$\text{ADVP} \rightarrow \text{ADVP CC ADVP}$	<i>now and two years ago</i>	[9.74]
$\text{ADJP} \rightarrow \text{ADJP CC ADJP}$	<i>quaint and rather deceptive</i>	[9.75]
$\text{SBAR} \rightarrow \text{SBAR CC SBAR}$	<i>whether they want control</i>	[9.76]
	<i>or whether they want exports</i>	

5176 9.2.4 Grammatical ambiguity

5177 Context-free parsing is useful not only because it determines whether a sentence is grammatical, but mainly because the constituents and their relations can be applied to tasks such as information extraction (chapter 17) and sentence compression (Jing, 2000; Clarke and Lapata, 2008). However, the **ambiguity** of wide-coverage natural language grammars poses a serious problem for such potential applications. As an example, Figure 9.13 shows 5182 two possible analyses for the simple sentence *We eat sushi with chopsticks*, depending on 5183 whether the *chopsticks* modify *eat* or *sushi*. Realistic grammars can license thousands or 5184 even millions of parses for individual sentences. **Weighted context-free grammars** solve 5185 this problem by attaching weights to each production, and selecting the derivation with 5186 the highest score. This is the focus of chapter 10.

5187 9.3 *Mildly context-sensitive languages

5188 Beyond context-free languages lie **context-sensitive languages**, in which the expansion 5189 of a non-terminal depends on its neighbors. In the general class of context-sensitive 5190 languages, computation becomes much more challenging: the membership problem for 5191 context-sensitive languages is PSPACE-complete. Since PSPACE contains the complexity 5192 class NP (problems that can be solved in polynomial time on a non-deterministic Turing

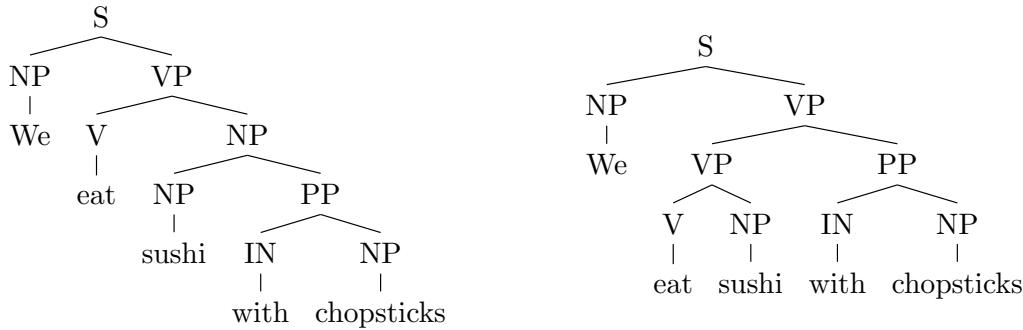


Figure 9.13: Two derivations of the same sentence

5193 machine), PSPACE-complete problems cannot be solved efficiently if $P \neq NP$. Thus, de-
 5194 signing an efficient parsing algorithm for the full class of context-sensitive languages is
 5195 probably hopeless.¹⁰

5196 However, Joshi (1985) identifies a set of properties that define **mildly context-sensitive**
 5197 **languages**, which are a strict subset of context-sensitive languages. Like context-free lan-
 5198 guages, mildly context-sensitive languages are parseable in polynomial time. However,
 5199 the mildly context-sensitive languages include non-context-free languages, such as the
 5200 “copy language” $\{ww \mid w \in \Sigma^*\}$ and the language $a^m b^n c^m d^n$. Both are characterized by
 5201 **cross-serial dependencies**, linking symbols at long distance across the string.¹¹ For exam-
 5202 ple, in the language $a^n b^m c^n d^m$, each a symbol is linked to exactly one c symbol, regardless
 5203 of the number of intervening b symbols.

5204 9.3.1 Context-sensitive phenomena in natural language

5205 Such phenomena are occasionally relevant to natural language. A classic example is found
 5206 in Swiss-German (Shieber, 1985), in which sentences such as *we let the children help Hans*
 5207 *paint the house* are realized by listing all nouns before all verbs, i.e., *we the children Hans the*
 5208 *house let help paint*. Furthermore, each noun’s determiner is dictated by the noun’s **case**
 5209 **marking** (the role it plays with respect to the verb). Using an argument that is analogous
 5210 to the earlier discussion of center-embedding (§ 9.2), Shieber describes these case marking
 5211 constraints as a set of cross-serial dependencies, homomorphic to $a^m b^n c^m d^n$, and therefore
 5212 not context-free.

¹⁰If $P \neq NP$, then it contains problems that cannot be solved in polynomial time on a non-deterministic Turing machine; equivalently, solutions to these problems cannot even be checked in polynomial time (Arora and Barak, 2009).

¹¹A further condition of the set of mildly-context-sensitive languages is *constant growth*: if the strings in the language are arranged by length, the gap in length between any pair of adjacent strings is bounded by some language specific constant. This condition excludes languages such as $\{a^{2^n} \mid n \geq 0\}$.

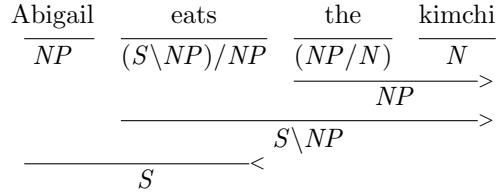


Figure 9.14: A syntactic analysis in CCG involving forward and backward function application

As with the move from regular to context-free languages, mildly context-sensitive languages can also be motivated by expedience. While finite sequences of cross-serial dependencies can in principle be handled in a context-free grammar, it is often more convenient to use a mildly context-sensitive formalism like **tree-adjoining grammar** (TAG) and **combinatory categorial grammar** (CCG). TAG-inspired parsers have been shown to be particularly effective in parsing the Penn Treebank (Collins, 1997; Carreras et al., 2008), and CCG plays a leading role in current research on semantic parsing (Zettlemoyer and Collins, 2005). These two formalisms are weakly equivalent: any language that can be specified in TAG can also be specified in CCG, and vice versa (Joshi et al., 1991). The remainder of the chapter gives a brief overview of CCG, but you are encouraged to consult Joshi and Schabes (1997) and Steedman and Baldridge (2011) for more detail on TAG and CCG respectively.

9.3.2 Combinatory categorial grammar

In combinatory categorial grammar, structural analyses are built up through a small set of generic combinatorial operations, which apply to immediately adjacent sub-structures. These operations act on the categories of the sub-structures, producing a new structure with a new category. The basic categories include S (sentence), NP (noun phrase), VP (verb phrase) and N (noun). The goal is to label the entire span of text as a sentence, S .

Complex categories, or types, are constructed from the basic categories, parentheses, and forward and backward slashes: for example, S/NP is a complex type, indicating a sentence that is lacking a noun phrase to its right; $S\backslash NP$ is a sentence lacking a noun phrase to its left. Complex types act as functions, and the most basic combinatory operations are function application to either the right or left neighbor. For example, the type of a verb phrase, such as *eats*, would be $S\backslash NP$. Applying this function to a subject noun phrase to its left results in an analysis of *Abigail eats* as category S , indicating a successful parse.

Transitive verbs must first be applied to the direct object, which in English appears to the right of the verb, before the subject, which appears on the left. They therefore have the more complex type $(S\backslash NP)/NP$. Similarly, the application of a determiner to the noun at

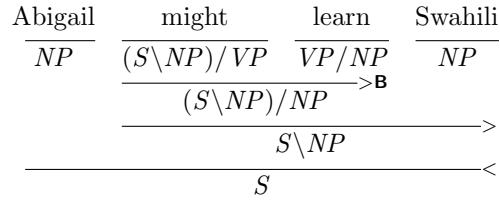


Figure 9.15: A syntactic analysis in CCG involving function composition (example modified from Steedman and Baldridge, 2011)

5242 its right results in a noun phrase, so determiners have the type NP/N. Figure 9.14 pro-
 5243 vides an example involving a transitive verb and a determiner. A key point from this
 5244 example is that it can be trivially transformed into phrase-structure tree, by treating each
 5245 function application as a constituent phrase. Indeed, when CCG's only combinatory op-
 5246 erators are forward and backward function application, it is equivalent to context-free
 5247 grammar. However, the location of the "effort" has changed. Rather than designing good
 5248 productions, the grammar designer must focus on the **lexicon** — choosing the right cate-
 5249 gories for each word. This makes it possible to parse a wide range of sentences using only
 5250 a few generic combinatory operators.

5251 Things become more interesting with the introduction of two additional operators:
 5252 **composition** and **type-raising**. Function composition enables the combination of com-
 5253 plex types: $X/Y \circ Y/Z \Rightarrow_B X/Z$ (forward composition) and $Y \setminus Z \circ X \setminus Y \Rightarrow_B X \setminus Z$ (back-
 5254 ward composition).¹² Composition makes it possible to "look inside" complex types, and
 5255 combine two adjacent units if the "input" for one is the "output" for the other. Figure 9.15
 5256 shows how function composition can be used to handle modal verbs. While this sen-
 5257 tence can be parsed using only function application, the composition-based analysis is
 5258 preferable because the unit *might learn* functions just like a transitive verb, as in the exam-
 5259 ple *Abigail studies Swahili*. This in turn makes it possible to analyze conjunctions such as
 5260 *Abigail studies and might learn Swahili*, attaching the direct object *Swahili* to the entire con-
 5261 joined verb phrase *studies and might learn*. The Penn Treebank grammar fragment from
 5262 § 9.2.3 would be unable to handle this case correctly: the direct object *Swahili* could attach
 5263 only to the second verb *learn*.

5264 Type raising converts an element of type X to a more complex type: $X \Rightarrow_T T/(T \setminus X)$
 5265 (forward type-raising to type T), and $X \Rightarrow_T T \setminus (T/X)$ (backward type-raising to type
 5266 T). Type-raising makes it possible to reverse the relationship between a function and its
 5267 argument — by transforming the argument into a function over functions over arguments!
 5268 An example may help. Figure 9.15 shows how to analyze an object relative clause, *a story*
 5269 *that Abigail tells*. The problem is that *tells* is a transitive verb, expecting a direct object to
 5270 its right. As a result, *Abigail tells* is not a valid constituent. The issue is resolved by raising

¹²The subscript **B** follows notation from Curry and Feys (1958).

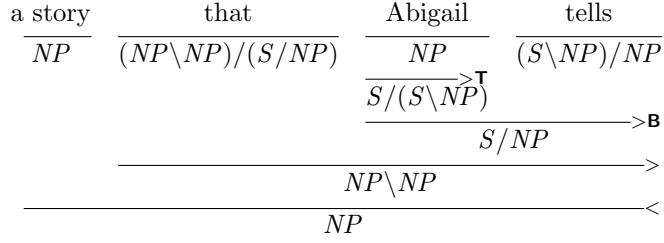


Figure 9.16: A syntactic analysis in CCG involving an object relative clause

5271 *Abigail* from NP to the complex type $(S/NP) \setminus NP$. This function can then be combined
 5272 with the transitive verb *tells* by forward composition, resulting in the type (S/NP) , which
 5273 is a sentence lacking a direct object to its right.¹³ From here, we need only design the
 5274 lexical entry for the complementizer *that* to expect a right neighbor of type (S/NP) , and
 5275 the remainder of the derivation can proceed by function application.

5276 Composition and type-raising give CCG considerable power and flexibility, but at a
 5277 price. The simple sentence *Abigail tells Max* can be parsed in two different ways: by func-
 5278 tion application (first forming the verb phrase *tells Max*), and by type-raising and compo-
 5279 sition (first forming the non-constituent *Abigail tells*). This **derivational ambiguity** does
 5280 not affect the resulting linguistic analysis, so it is sometimes known as **spurious ambi-**
 5281 **guity**. Hockenmaier and Steedman (2007) present a translation algorithm for converting
 5282 the Penn Treebank into CCG derivations, using composition and type-raising only when
 5283 necessary.

5284 Exercises

- 5285 1. Sketch out the state diagram for finite-state acceptors for the following languages
 5286 on the alphabet $\{a, b\}$.
 - 5287 a) Even-length strings. (Be sure to include 0 as an even number.)
 - 5288 b) Strings that contain *aaa* as a substring.
 - 5289 c) Strings containing an even number of *a* and an odd number of *b* symbols.
 - 5290 d) Strings in which the substring *bbb* must be terminal if it appears — the string
 5291 need not contain *bbb*, but if it does, nothing can come after it.
- 5292 2. Levenshtein edit distance is the number of insertions, substitutions, or deletions
 5293 required to convert one string to another.

¹³The missing direct object would be analyzed as a **trace** in CFG-like approaches to syntax, including the Penn Treebank.

- 5294 a) Define a finite-state acceptor that accepts all strings with edit distance 1 from
 5295 the target string, *target*.
 5296 b) Now think about how to generalize your design to accept all strings with edit
 5297 distance from the target string equal to d . If the target string has length ℓ , what
 5298 is the minimal number of states required?
- 5299 3. Construct an FSA in the style of Figure 9.3, which handles the following examples:
- 5300 • *nation*/N, *national*/ADJ, *nationalize*/V, *nationalizer*/N
 5301 • *America*/N, *American*/ADJ, *Americanize*/V, *Americanizer*/N
- 5302 Be sure that your FSA does not accept any further derivations, such as **nationalizeral*
 5303 and **Americanizern*.
- 5304 4. Show how to construct a trigram language model in a weighted finite-state acceptor.
 5305 Make sure that you handle the edge cases at the beginning and end of the input.
- 5306 5. Extend the FST in Figure 9.6 to handle the other two parts of rule 1a of the Porter
 5307 stemmer: *-sses* → *ss*, and *-ies* → *-i*.
- 5308 6. § 9.1.4 describes T_O , a transducer that captures English orthography by transduc-
 5309 ing *cook + ed* → *cooked* and *bake + ed* → *baked*. Design an unweighted finite-state
 5310 transducer that captures this property of English orthography.
 5311 Next, augment the transducer to appropriately model the suffix *-s* when applied to
 5312 words ending in *s*, e.g. *kiss+s* → *kisses*.
- 5313 7. Add parenthesization to the grammar in Figure 9.11 so that it is no longer ambigu-
 5314 ous.
- 5315 8. Construct three examples — a noun phrase, a verb phrase, and a sentence — which
 5316 can be derived from the Penn Treebank grammar fragment in § 9.2.3, yet are not
 5317 grammatical. Avoid reusing examples from the text. Optionally, propose corrections
 5318 to the grammar to avoid generating these cases.
- 5319 9. Produce parses for the following sentences, using the Penn Treebank grammar frag-
 5320 ment from § 9.2.3.
- 5321 (9.8) This aggression will not stand.
 5322 (9.9) I can get you a toe.
 5323 (9.10) Sometimes you eat the bar and sometimes the bar eats you.

5324 Then produce parses for three short sentences from a news article from this week.

5325 10. * One advantage of CCG is its flexibility in handling coordination:

5326 (9.11) a. *Hunter and Tristan speak Hawaiian*

5327 b. *Hunter speaks and Tristan understands Hawaiian*

Define the lexical entry for *and* as

$$\textit{and} := (X/X) \setminus X, \quad [9.77]$$

5328 where X can refer to any type. Using this lexical entry, show how to parse the two
5329 examples above. In the second example, *Swahili* should be combined with the coor-
5330 dination *Abigail speaks and Max understands*, and not just with the verb *understands*.

5331 **Chapter 10**

5332 **Context-free parsing**

5333 Parsing is the task of determining whether a string can be derived from a given context-
5334 free grammar, and if so, how. A parser’s output is a tree, like the ones shown in Fig-
5335 ure 9.13. Such trees can answer basic questions of who-did-what-to-whom, and have ap-
5336 plications in downstream tasks like semantic analysis (chapter 12 and 13) and information
5337 extraction (chapter 17).

For a given input and grammar, how many parse trees are there? Consider a minimal context-free grammar with only one non-terminal, X , and the following productions:

$$\begin{aligned} X &\rightarrow X \ X \\ X &\rightarrow aardvark \mid abacus \mid \dots \mid zyther \end{aligned}$$

The second line indicates unary productions to every nonterminal in Σ . In this grammar, the number of possible derivations for a string w is equal to the number of binary bracketings, e.g.,

$$(((w_1 w_2) w_3) w_4) w_5), \quad (((w_1 (w_2 w_3)) w_4) w_5), \quad ((w_1 (w_2 (w_3 w_4))) w_5), \quad \dots$$

5338 The number of such bracketings is a **Catalan number**, which grows super-exponentially
5339 in the length of the sentence, $C_n = \frac{(2n)!}{(n+1)n!}$. As with sequence labeling, it is only possible to
5340 exhaustively search the space of parses by resorting to locality assumptions, which make it
5341 possible to search efficiently by reusing shared substructures with dynamic programming.
5342 This chapter focuses on a bottom-up dynamic programming algorithm, which enables
5343 exhaustive search of the space of possible parses, but imposes strict limitations on the
5344 form of scoring function. These limitations can be relaxed by abandoning exhaustive
5345 search. Non-exact search methods will be briefly discussed at the end of this chapter, and
5346 one of them — **transition-based parsing** — will be the focus of chapter 11.

S	\rightarrow	NP VP
NP	\rightarrow	NP PP we sushi chopsticks
PP	\rightarrow	IN NP
IN	\rightarrow	with
VP	\rightarrow	V NP VP PP
V	\rightarrow	eat

Table 10.1: A toy example context-free grammar

5347 **10.1 Deterministic bottom-up parsing**

5348 The **CKY algorithm**¹ is a bottom-up approach to parsing in a context-free grammar. It
 5349 efficiently tests whether a string is in a language, without enumerating all possible parses.
 5350 The algorithm first forms small constituents, and then tries to merge them into larger
 5351 constituents.

5352 To understand the algorithm, consider the input, *We eat sushi with chopsticks*. According-
 5353 ing to the toy grammar in Table 10.1, each terminal symbol can be generated by exactly
 5354 one unary production, resulting in the sequence NP V NP IN NP. In real examples, there
 5355 may be many unary productions for each individual token. In any case, the next step
 5356 is to try to apply binary productions to merge adjacent symbols into larger constituents:
 5357 for example, V NP can be merged into a verb phrase (VP), and IN NP can be merged
 5358 into a prepositional phrase (PP). Bottom-up parsing searches for a series of mergers that
 5359 ultimately results in the start symbol S covering the entire input.

5360 The CKY algorithm systematizes this search by incrementally constructing a table t in
 5361 which each cell $t[i, j]$ contains the set of nonterminals that can derive the span $w_{i+1:j}$. The
 5362 algorithm fills in the upper right triangle of the table; it begins with the diagonal, which
 5363 corresponds to substrings of length 1, and then computes derivations for progressively
 5364 larger substrings, until reaching the upper right corner $t[0, M]$, which corresponds to the
 5365 entire input, $w_{1:M}$. If the start symbol S is in $t[0, M]$, then the string w is in the language
 5366 defined by the grammar. This process is detailed in Algorithm 13, and the resulting data
 5367 structure is shown in Figure 10.1. Informally, here's how it works:

- 5368 • Begin by filling in the diagonal: the cells $t[m - 1, m]$ for all $m \in \{1, 2, \dots, M\}$. These
 5369 cells are filled with terminal productions that yield the individual tokens; for the
 5370 word $w_2 = \text{sushi}$, we fill in $t[1, 2] = \{\text{NP}\}$, and so on.
- 5371 • Then fill in the next diagonal, in which each cell corresponds to a subsequence of
 5372 length two: $t[0, 2], t[1, 3], \dots, t[M - 2, M]$. These cells are filled in by looking for

¹The name is for Cocke-Kasami-Younger, the inventors of the algorithm. It is a special case of **chart parsing**, because its stores reusable computations in a chart-like data structure.

binary productions capable of producing at least one entry in each of the cells corresponding to left and right children. For example, VP can be placed in the cell $t[1, 3]$ because the grammar includes the production $VP \rightarrow V\ NP$, and because the chart contains $V \in t[1, 2]$ and $NP \in t[2, 3]$.

- At the next diagonal, the entries correspond to spans of length three. At this level, there is an additional decision at each cell: where to split the left and right children. The cell $t[i, j]$ corresponds to the subsequence $w_{i+1:j}$, and we must choose some *split point* $i < k < j$, so that the span $w_{i+1:k}$ is the left child, and the span $w_{k+1:j}$ is the right child. We consider all possible k , looking for productions that generate elements in $t[i, k]$ and $t[k, j]$; the left-hand side of all such productions can be added to $t[i, j]$. When it is time to compute $t[i, j]$, the cells $t[i, k]$ and $t[k, j]$ are guaranteed to be complete, since these cells correspond to shorter sub-strings of the input.
- The process continues until we reach $t[0, M]$.

Figure 10.1 shows the chart that arises from parsing the sentence *We eat sushi with chopsticks* using the grammar defined above.

10.1.1 Recovering the parse tree

As with the Viterbi algorithm, it is possible to identify a successful parse by storing and traversing an additional table of back-pointers. If we add an entry X to cell $t[i, j]$ by using the production $X \rightarrow YZ$ and the split point k , then we store the back-pointer $b[i, j, X] = (Y, Z, k)$. Once the table is complete, we can recover a parse by tracing these pointers, starting at $b[0, M, S]$, and stopping when they ground out at terminal productions.

For ambiguous sentences, there will be multiple paths to reach $S \in t[0, M]$. For example, in Figure 10.1, the goal state $S \in t[0, M]$ is reached through the state $VP \in t[1, 5]$, and there are two different ways to generate this constituent: one with *(eat sushi)* and *(with chopsticks)* as children, and another with *(eat)* and *(sushi with chopsticks)* as children. The presence of multiple paths indicates that the input can be generated by the grammar in more than one way. In Algorithm 13, one of these derivations is selected arbitrarily. As discussed in § 10.3, **weighted context-free grammars** compute a score for all permissible derivations, and a minor modification of CKY allows it to identify the single derivation with the maximum score.

10.1.2 Non-binary productions

As presented above, the CKY algorithm assumes that all productions with non-terminals on the right-hand side (RHS) are binary. In real grammars, such as the one considered in chapter 9, there are other types of productions: some have more than two elements on the right-hand side, and others produce a single non-terminal.

Algorithm 13 The CKY algorithm for parsing a sequence $w \in \Sigma^*$ in a context-free grammar $G = (N, \Sigma, R, S)$, with non-terminals N , production rules R , and start symbol S . The grammar is assumed to be in Chomsky normal form (§ 9.2.1). The function $\text{PICKFROM}(b[i, j, X])$ selects an element of the set $b[i, j, X]$ arbitrarily. All values of t and b are initialized to \emptyset .

```

1: procedure CKY( $w, G = (N, \Sigma, R, S)$ )
2:   for  $m \in \{1 \dots M\}$  do
3:      $t[m - 1, m] \leftarrow \{X : (X \rightarrow w_m) \in R\}$ 
4:   for  $\ell \in \{2, 3, \dots, M\}$  do                                 $\triangleright$  Iterate over constituent lengths
5:     for  $m \in \{0, 1, \dots, M - \ell\}$  do                       $\triangleright$  Iterate over left endpoints
6:       for  $k \in \{m + 1, m + 2, \dots, m + \ell - 1\}$  do           $\triangleright$  Iterate over split points
7:         for  $(X \rightarrow Y Z) \in R$  do                       $\triangleright$  Iterate over rules
8:           if  $Y \in t[m, k] \wedge Z \in t[k, m + \ell]$  then
9:              $t[m, m + \ell] \leftarrow t[m, m + \ell] \cup X$        $\triangleright$  Add non-terminal to table
10:             $b[m, m + \ell, X] \leftarrow b[m, m + \ell, X] \cup (Y, Z, k)$      $\triangleright$  Add back-pointers
11:   if  $S \in t[0, M]$  then
12:     return TRACEBACK( $S, 0, M, b$ )
13:   else
14:     return  $\emptyset$ 
15: procedure TRACEBACK( $X, i, j, b$ )
16:   if  $j = i + 1$  then
17:     return  $X$ 
18:   else
19:      $(Y, Z, k) \leftarrow \text{PICKFROM}(b[i, j, X])$ 
20:     return  $X \rightarrow (\text{TRACEBACK}(Y, i, k, b), \text{TRACEBACK}(Z, k, j, b))$ 

```

- 5408 • Productions with more than two elements on the right-hand side can be **binarized**
 5409 by creating additional non-terminals, as described in § 9.2.1. For example, the pro-
 5410 duction $VP \rightarrow V NP NP$ (for ditransitive verbs) can be converted to $VP \rightarrow VP_{ditrans}/NP NP$,
 5411 by adding the non-terminal $VP_{ditrans}/NP$ and the production $VP_{ditrans}/NP \rightarrow V NP$.
- 5412 • What about unary productions like $VP \rightarrow V$? While such productions are not a
 5413 part of Chomsky Normal Form — and can therefore be eliminated in preprocessing
 5414 the grammar — in practice, a more typical solution is to modify the CKY algorithm.
 5415 The algorithm makes a second pass on each diagonal in the table, augmenting each
 5416 cell $t[i, j]$ with all possible unary productions capable of generating each item al-
 5417 ready in the cell: formally, $t[i, j]$ is extended to its **unary closure**. Suppose the ex-
 5418 ample grammar in Table 10.1 were extended to include the production $VP \rightarrow V$,
 5419 enabling sentences with intransitive verb phrases, like *we eat*. Then the cell $t[1, 2]$
 5420 — corresponding to the word *eat* — would first include the set $\{V\}$, and would be
 5421 augmented to the set $\{V, VP\}$ during this second pass.

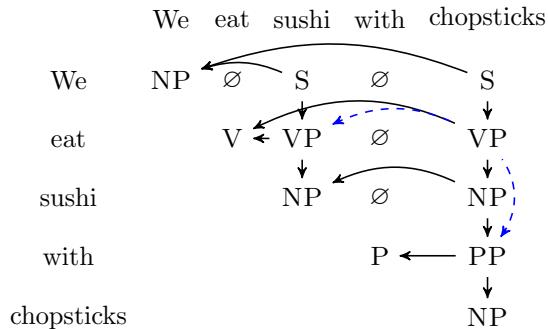


Figure 10.1: An example completed CKY chart. The solid and dashed lines show the back pointers resulting from the two different derivations of VP in position $t[1, 5]$.

5422 10.1.3 Complexity

5423 For an input of length M and a grammar with R productions and N non-terminals, the
 5424 space complexity of the CKY algorithm is $\mathcal{O}(M^2N)$: the number of cells in the chart is
 5425 $\mathcal{O}(M^2)$, and each cell must hold $\mathcal{O}(N)$ elements. The time complexity is $\mathcal{O}(M^3R)$: each
 5426 cell is computed by searching over $\mathcal{O}(M)$ split points, with R possible productions for
 5427 each split point. Both the time and space complexity are considerably worse than the
 5428 Viterbi algorithm, which is linear in the length of the input.

5429 10.2 Ambiguity

5430 In natural language, there is rarely a single parse for a given sentence. The main culprit is
 5431 ambiguity, which is endemic to natural language syntax. Here are a few broad categories:

- 5432 • **Attachment ambiguity:** e.g., *We eat sushi with chopsticks, I shot an elephant in my pajamas*. In these examples, the prepositions (*with, in*) can attach to either the verb
 5433 or the direct object.
- 5435 • **Modifier scope:** e.g., *southern food store, plastic cup holder*. In these examples, the first
 5436 word could be modifying the subsequent adjective, or the final noun.
- 5437 • **Particle versus preposition:** e.g., *The puppy tore up the staircase*. Phrasal verbs like
 5438 *tore up* often include particles which could also act as prepositions. This has struc-
 5439 tural implications: if *up* is a preposition, then *up the staircase* is a prepositional
 5440 phrase; if *up* is a particle, then *the staircase* is the direct object to the verb.
- 5441 • **Complement structure:** e.g., *The students complained to the professor that they didn't
 5442 understand*. This is another form of attachment ambiguity, where the complement

5443 *that they didn't understand* could attach to the main verb (*complained*), or to the indi-
 5444 rect object (*the professor*).

5445 • **Coordination scope:** e.g., “I see,” said the blind man, as he picked up the hammer and
 5446 saw. In this example, the lexical ambiguity for *saw* enables it to be coordinated either
 5447 with the noun *hammer* or the verb *picked up*.

5448 These forms of ambiguity can combine, so that seemingly simple headlines like *Fed*
 5449 *raises interest rates* have dozens of possible analyses even in a minimal grammar. In a
 5450 broad coverage grammar, typical sentences can have millions of parses. While careful
 5451 grammar design can chip away at this ambiguity, a better strategy is combine broad cov-
 5452 erage parsers with data-driven strategies for identifying the correct analysis.

5453 10.2.1 Parser evaluation

5454 Before continuing to parsing algorithms that are able to handle ambiguity, let us stop
 5455 to consider how to measure parsing performance. Suppose we have a set of *reference*
 5456 *parses* — the ground truth — and a set of *system parses* that we would like to score. A
 5457 simple solution would be per-sentence accuracy: the parser is scored by the proportion of
 5458 sentences on which the system and reference parses exactly match.² But as any student
 5459 knows, it always nice to get *partial credit*, which we can assign to analyses that correctly
 5460 match parts of the reference parse. The PARSEval metrics (Grishman et al., 1992) score
 5461 each system parse via:

5462 **Precision:** the fraction of constituents in the system parse that match a constituent in the
 5463 reference parse.

5464 **Recall:** the fraction of constituents in the reference parse that match a constituent in the
 5465 system parse.

5466 In **labeled precision** and **recall**, the system must also match the phrase type for each
 5467 constituent; in **unlabeled precision** and **recall**, it is only required to match the constituent
 5468 structure. As described in chapter 4, the precision and recall can be combined into an
 5469 F-MEASURE by their harmonic mean.

5470 Suppose that the left tree of Figure 10.2 is the system parse, and that the right tree is
 5471 the reference parse. Then:

5472 • $S \rightarrow w_{1:5}$ is *true positive*, because it appears in both trees.

²Most parsing papers do not report results on this metric, but Suzuki et al. (2018) find that a strong parser recovers the exact parse in roughly 50% of all sentences. Performance on short sentences is generally much higher.

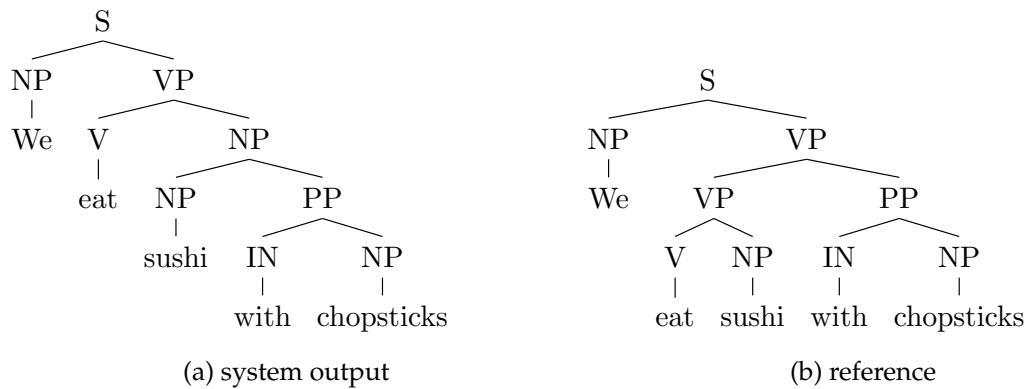


Figure 10.2: Two possible analyses from the grammar in Table 10.1

- VP → $w_{2:5}$ is *true positive* as well.
 - NP → $w_{3:5}$ is *false positive*, because it appears only in the system output.
 - PP → $w_{4:5}$ is *true positive*, because it appears in both trees.
 - VP → $w_{2:3}$ is *false negative*, because it appears only in the reference.

5477 The labeled and unlabeled precision of this parse is $\frac{3}{4} = 0.75$, and the recall is $\frac{3}{4} = 0.75$, for
 5478 an F-measure of 0.75. For an example in which precision and recall are not equal, suppose
 5479 the reference parse instead included the production $VP \rightarrow V NP PP$. In this parse, the
 5480 reference does not contain the constituent $w_{2,3}$, so the recall would be 1.³

5481 10.2.2 Local solutions

5482 Some ambiguity can be resolved locally. Consider the following examples,

- 5483 (10.1) a. We met the President on Monday.
5484 b. We met the President of Mexico.

Each case ends with a prepositional phrase, which can be attached to the verb *met* or the noun phrase *the president*. If given a labeled corpus, we can compare the likelihood of the observing the preposition alongside each candidate attachment point,

$$p(on \mid met) \geq p(on \mid President) \quad [10.1]$$

$$p(of \mid met) \geq p(of \mid President). \quad [10.2]$$

³While the grammar must be binarized before applying the CKY algorithm, evaluation is performed on the original parses. It is therefore necessary to “unbinarize” the output of a CKY-based parser, converting it back to the original grammar.

5485 A comparison of these probabilities would successfully resolve this case (Hindle and
 5486 Rooth, 1993). Other cases, such as the example *we eat sushi with chopsticks*, require con-
 5487 sidering the object of the preposition: consider the alternative *we eat sushi with soy sauce*.
 5488 With sufficient labeled data, some instances of attachment ambiguity can be solved by
 5489 supervised classification (Ratnaparkhi et al., 1994).

5490 However, there are inherent limitations to local solutions. While toy examples may
 5491 have just a few ambiguities to resolve, realistic sentences have thousands or millions of
 5492 possible parses. Furthermore, attachment decisions are interdependent, as shown in the
 5493 garden path example:

5494 (10.2) Cats scratch people with claws with knives.

5495 We may want to attach *with claws* to *scratch*, as would be correct in the shorter sentence
 5496 in *cats scratch people with claws*. But this leaves nowhere to attach *with knives*. The cor-
 5497 rect interpretation can be identified only by considering the attachment decisions jointly.
 5498 The huge number of potential parses may seem to make exhaustive search impossible.
 5499 But as with sequence labeling, locality assumptions make it possible to search this space
 5500 efficiently.

5501 10.3 Weighted Context-Free Grammars

5502 Let us define a derivation τ as a set of **anchored productions**,

$$\tau = \{X \rightarrow \alpha, (i, j, k)\}, \quad [10.3]$$

5503 with X corresponding to the left-hand side non-terminal and α corresponding to the right-
 5504 hand side. For grammars in Chomsky normal form, α is either a pair of non-terminals or
 5505 a terminal symbol. The indices i, j, k anchor the production in the input, with X deriving
 5506 the span $w_{i+1:j}$. For binary productions, $w_{i+1:k}$ indicates the span of the left child, and
 5507 $w_{k+1:j}$ indicates the span of the right child; for unary productions, k is ignored. For an
 5508 input w , the optimal parse is,

$$\hat{\tau} = \underset{\tau \in \mathcal{T}(w)}{\operatorname{argmax}} \Psi(\tau), \quad [10.4]$$

5509 where $\mathcal{T}(w)$ is the set of derivations that yield the input w .

5510 Define a scoring function Ψ that decomposes across anchored productions,

$$\Psi(\tau) = \sum_{(X \rightarrow \alpha, (i, j, k)) \in \tau} \psi(X \rightarrow \alpha, (i, j, k)). \quad [10.5]$$

5511 This is a locality assumption, akin to the assumption in Viterbi sequence labeling. In this
 5512 case, the assumption states that the overall score is a sum over scores of productions,

		$\psi(\cdot)$	$\exp \psi(\cdot)$
S	\rightarrow NP VP	0	1
NP	\rightarrow NP PP	-1	$\frac{1}{2}$
	\rightarrow <i>we</i>	-2	$\frac{1}{4}$
	\rightarrow <i>sushi</i>	-3	$\frac{1}{8}$
	\rightarrow <i>chopsticks</i>	-3	$\frac{1}{8}$
PP	\rightarrow IN NP	0	1
IN	\rightarrow <i>with</i>	0	1
VP	\rightarrow V NP	-1	$\frac{1}{2}$
	\rightarrow VP PP	-2	$\frac{1}{4}$
	\rightarrow MD V	-2	$\frac{1}{4}$
V	\rightarrow <i>eat</i>	0	1

Table 10.2: An example weighted context-free grammar (WCFG). The weights are chosen so that $\exp \psi(\cdot)$ sums to one over right-hand sides for each non-terminal; this is required by probabilistic context-free grammars, but not by WCFGs in general.

5513 which are computed independently. In a **weighted context-free grammar** (WCFG), the
 5514 score of each anchored production $X \rightarrow (\alpha, (i, j, k))$ is simply $\psi(X \rightarrow \alpha)$, ignoring the
 5515 anchor (i, j, k) . In other parsing models, the anchors can be used to access features of the
 5516 input, while still permitting efficient bottom-up parsing.

Example Consider the weighted grammar shown in Table 10.2, and the analysis in Figure 10.2b.

$$\begin{aligned} \Psi(\tau) = & \psi(S \rightarrow NP\ VP) + \psi(VP \rightarrow VP\ PP) + \psi(VP \rightarrow V\ NP) + \psi(PP \rightarrow IN\ NP) \\ & + \psi(NP \rightarrow We) + \psi(V \rightarrow eat) + \psi(NP \rightarrow sushi) + \psi(IN \rightarrow with) + \psi(NP \rightarrow chopsticks) \end{aligned} \quad [10.6]$$

$$= 0 - 2 - 1 + 0 - 2 + 0 - 3 + 0 - 3 = -11. \quad [10.7]$$

5517 In the alternative parse in Figure 10.2a, the production $VP \rightarrow VP\ PP$ (with score -2) is
 5518 replaced with the production $NP \rightarrow NP\ PP$ (with score -1); all other productions are the
 5519 same. As a result, the score for this parse is -10. This example hints at a problem with
 5520 WCFG parsing on non-terminals such as NP, VP, and PP: a WCFG will *always* prefer
 5521 either VP or NP attachment, regardless of what is being attached! Solutions to this issue
 5522 are discussed in § 10.5.

Algorithm 14 CKY algorithm for parsing a string $w \in \Sigma^*$ in a weighted context-free grammar (N, Σ, R, S) , where N is the set of non-terminals and R is the set of weighted productions. The grammar is assumed to be in Chomsky normal form (§ 9.2.1). The function TRACEBACK is defined in Algorithm 13.

```

procedure WCKY( $w, G = (N, \Sigma, R, S)$ )
  for all  $i, j, X$  do ▷ Initialization
     $t[i, j, X] \leftarrow 0$ 
     $b[i, j, X] \leftarrow \emptyset$ 
  for  $m \in \{1, 2, \dots, M\}$  do
    for all  $X \in N$  do
       $t[m, m + 1, X] \leftarrow \psi(X \rightarrow w_m, (m, m + 1, m))$ 
  for  $\ell \in \{2, 3, \dots, M\}$  do
    for  $m \in \{0, 1, \dots, M - \ell\}$  do
      for  $k \in \{m + 1, m + 2, \dots, m + \ell - 1\}$  do
         $t[m, m + \ell, X] \leftarrow \max_{k, Y, Z} \psi(X \rightarrow Y Z, (m, m + \ell, k)) + t[m, k, Y] + t[k, m + \ell, Z]$ 
         $b[m, m + \ell, X] \leftarrow \operatorname{argmax}_{k, Y, Z} \psi(X \rightarrow Y Z, (m + \ell, k)) + t[m, k, Y] + t[k, m + \ell, Z]$ 
  return TRACEBACK( $S, 0, M, b$ )

```

5523 **10.3.1 Parsing with weighted context-free grammars**

5524 The optimization problem in Equation 10.4 can be solved by modifying the CKY algo-
 5525 rithm. In the deterministic CKY algorithm, each cell $t[i, j]$ stored a set of non-terminals
 5526 capable of deriving the span $w_{i+1:j}$. We now augment the table so that the cell $t[i, j, X]$
 5527 is the *score of the best derivation* of $w_{i+1:j}$ from non-terminal X . This score is computed
 5528 recursively: for the anchored binary production $(X \rightarrow Y Z, (i, j, k))$, we compute:

- 5529 • the score of the anchored production, $\psi(X \rightarrow Y Z, (i, j, k))$;
- 5530 • the score of the best derivation of the left child, $t[i, k, Y]$;
- 5531 • the score of the best derivation of the right child, $t[k, j, Z]$.

5532 These scores are combined by addition. As in the unweighted CKY algorithm, the table
 5533 is constructed by considering spans of increasing length, so the scores for spans $t[i, k, Y]$
 5534 and $t[k, j, Z]$ are guaranteed to be available at the time we compute the score $t[i, j, X]$. The
 5535 value $t[0, M, S]$ is the score of the best derivation of w from the grammar. Algorithm 14
 5536 formalizes this procedure.

5537 As in unweighted CKY, the parse is recovered from the table of back pointers b , where
 5538 each $b[i, j, X]$ stores the argmax split point k and production $X \rightarrow Y Z$ in the derivation of
 5539 $w_{i+1:j}$ from X . The top scoring parse can be obtained by tracing these pointers backwards
 5540 from $b[0, M, S]$, all the way to the terminal symbols. This is analogous to the computation

of the best sequence of labels in the Viterbi algorithm by tracing pointers backwards from the end of the trellis. Note that we need only store back-pointers for the *best* path to $t[i, j, X]$; this follows from the locality assumption that the global score for a parse is a combination of the local scores of each production in the parse.

Example Let's revisit the parsing table in Figure 10.1. In a weighted CFG, each cell would include a score for each non-terminal; non-terminals that cannot be generated are assumed to have a score of $-\infty$. The first diagonal contains the scores of unary productions: $t[0, 1, \text{NP}] = -2$, $t[1, 2, \text{V}] = 0$, and so on. The next diagonal contains the scores for spans of length 2: $t[1, 3, \text{VP}] = -1 + 0 - 3 = -4$, $t[3, 5, \text{PP}] = 0 + 0 - 3 = -3$, and so on. Things get interesting when we reach the cell $t[1, 5, \text{VP}]$, which contains the score for the derivation of the span $w_{2:5}$ from the non-terminal VP. This score is computed as a max over two alternatives,

$$t[1, 5, \text{VP}] = \max(\psi(\text{VP} \rightarrow \text{VP PP}, (1, 3, 5)) + t[1, 3, \text{VP}] + t[3, 5, \text{PP}], \psi(\text{VP} \rightarrow \text{V NP}, (1, 2, 5)) + t[1, 2, \text{V}] + t[2, 5, \text{NP}]) \quad [10.8]$$

$$= \max(-2 - 4 - 3, -1 + 0 - 7) = -8. \quad [10.9]$$

Since the second case is the argmax, we set the back-pointer $b[1, 5, \text{VP}] = (\text{V}, \text{NP}, 2)$, enabling the optimal derivation to be recovered.

10.3.2 Probabilistic context-free grammars

Probabilistic context-free grammars (PCFGs) are a special case of weighted context-free grammars that arises when the weights correspond to probabilities. Specifically, the weight $\psi(X \rightarrow \alpha, (i, j, k)) = \log p(\alpha | X)$, where the probability of the right-hand side α is conditioned on the non-terminal X , and the anchor (i, j, k) is ignored. These probabilities must be normalized over all possible right-hand sides, so that $\sum_\alpha p(\alpha | X) = 1$, for all X . For a given parse τ , the product of the probabilities of the productions is equal to $p(\tau)$, under the **generative model** $\tau \sim \text{DRAWSUBTREE}(S)$, where the function `DRAWSUBTREE` is defined in Algorithm 15.

The conditional probability of a parse given a string is,

$$p(\tau | w) = \frac{p(\tau)}{\sum_{\tau' \in \mathcal{T}(w)} p(\tau')} = \frac{\exp \Psi(\tau)}{\sum_{\tau' \in \mathcal{T}(w)} \exp \Psi(\tau')}, \quad [10.10]$$

where $\Psi(\tau) = \sum_{X \rightarrow \alpha, (i, j, k) \in \tau} \psi(X \rightarrow \alpha, (i, j, k))$. Because the probability is monotonic in the score $\Psi(\tau)$, the maximum likelihood parse can be identified by the CKY algorithm without modification. If a normalized probability $p(\tau | w)$ is required, the denominator of Equation 10.10 can be computed by the **inside recurrence**, described below.

Algorithm 15 Generative model for derivations from probabilistic context-free grammars in Chomsky Normal Form (CNF).

```

procedure DRAWSUBTREE( $X$ )
    sample  $(X \rightarrow \alpha) \sim p(\alpha | X)$ 
    if  $\alpha = (Y Z)$  then
        return DRAWSUBTREE( $Y$ )  $\cup$  DRAWSUBTREE( $Z$ )
    else
        return  $(X \rightarrow \alpha)$             $\triangleright$  In CNF, all unary productions yield terminal symbols

```

Example The WCFG in Table 10.2 is designed so that the weights are log-probabilities, satisfying the constraint $\sum_{\alpha} \exp \psi(X \rightarrow \alpha) = 1$. As noted earlier, there are two parses in \mathcal{T} (*we eat sushi with chopsticks*), with scores $\Psi(\tau_1) = \log p(\tau_1) = -10$ and $\Psi(\tau_2) = \log p(\tau_2) = -11$. Therefore, the conditional probability $p(\tau_1 | w)$ is equal to,

$$p(\tau_1 | w) = \frac{p(\tau_1)}{p(\tau_1) + p(\tau_2)} = \frac{\exp \Psi(\tau_1)}{\exp \Psi(\tau_1) + \exp \Psi(\tau_2)} = \frac{2^{-10}}{2^{-10} + 2^{-11}} = \frac{2}{3}. \quad [10.11]$$

5561 **The inside recurrence** The denominator of Equation 10.10 can be viewed as a language
5562 model, summing over all valid derivations of the string w ,

$$p(w) = \sum_{\tau': \text{yield}(\tau')=w} p(\tau'). \quad [10.12]$$

Just as the CKY algorithm makes it possible to maximize over all such analyses, with a few modifications it can also compute their sum. Each cell $t[i, j, X]$ must store the log probability of deriving $w_{i+1:j}$ from non-terminal X . To compute this, we replace the maximization over split points k and productions $X \rightarrow Y Z$ with a “log-sum-exp” operation, which exponentiates the log probabilities of the production and the children, sums them in probability space, and then converts back to the log domain:

$$t[i, j, X] = \log \sum_{k, Y, Z} \exp (\psi(X \rightarrow Y Z) + t[i, k, Y] + t[k, j, Z]) \quad [10.13]$$

$$= \log \sum_{k, Y, Z} \exp (\log p(Y Z | X) + \log p(Y \rightarrow w_{i+1:k}) + \log p(Z \rightarrow w_{k+1:j})) \quad [10.14]$$

$$= \log \sum_{k, Y, Z} p(Y Z | X) \times p(Y \rightarrow w_{i+1:k}) \times p(Z \rightarrow w_{k+1:j}) \quad [10.15]$$

$$= \log \sum_{k, Y, Z} p(Y Z, w_{i+1:k}, w_{k+1:j} | X) \quad [10.16]$$

$$= \log p(X \rightsquigarrow w_{i+1:j}), \quad [10.17]$$

5563 with $X \rightsquigarrow w_{i+1:j}$ indicating the event that non-terminal X yields the span $w_{i+1}, w_{i+2}, \dots, w_j$.
 5564 The recursive computation of $t[i, j, X]$ is called the **inside recurrence**, because it computes
 5565 the probability of each subtree as a combination of the probabilities of the smaller subtrees
 5566 that are inside of it. The name implies a corresponding **outside recurrence**, which com-
 5567 putes the probability of a non-terminal X spanning $w_{i+1:j}$, joint with the outside context
 5568 ($w_{1:i}, w_{j+1:M}$). This recurrence is described in § 10.4.3. The inside and outside recurrences
 5569 are analogous to the forward and backward recurrences in probabilistic sequence label-
 5570 ing (see § 7.5.3). They can be used to compute the marginal probabilities of individual
 5571 anchored productions, $p(X \rightarrow \alpha, (i, j, k) \mid \mathbf{w})$, summing over all possible derivations of
 5572 \mathbf{w} .

5573 **10.3.3 *Semiring weighted context-free grammars**

The weighted and unweighted CKY algorithms can be unified with the inside recurrence using the same semiring notation described in § 7.7.3. The generalized recurrence is:

$$t[i, j, X] = \bigoplus_{k, Y, Z} \psi(X \rightarrow Y Z, (i, j, k)) \otimes t[i, k, Y] \otimes t[k, j, Z]. \quad [10.18]$$

5574 This recurrence subsumes all of the algorithms that have been discussed in this chapter to
 5575 this point.

5576 **Unweighted CKY.** When $\psi(X \rightarrow \alpha, (i, j, k))$ is a *Boolean truth value* $\{\top, \perp\}$, \otimes is logical
 5577 conjunction, and \bigoplus is logical disjunction, then we derive CKY recurrence for un-
 5578 weighted context-free grammars, discussed in § 10.1 and Algorithm 13.

5579 **Weighted CKY.** When $\psi(X \rightarrow \alpha, (i, j, k))$ is a scalar score, \otimes is addition, and \bigoplus is maxi-
 5580 mization, then we derive the CKY recurrence for weighted context-free grammars,
 5581 discussed in § 10.3 and Algorithm 14. When $\psi(X \rightarrow \alpha, (i, j, k)) = \log p(\alpha \mid X)$,
 5582 this same setting derives the CKY recurrence for finding the maximum likelihood
 5583 derivation in a probabilistic context-free grammar.

5584 **Inside recurrence.** When $\psi(X \rightarrow \alpha, (i, j, k))$ is a log probability, \otimes is addition, and $\bigoplus =$
 5585 $\log \sum \exp$, then we derive the inside recurrence for probabilistic context-free gram-
 5586 mmars, discussed in § 10.3.2. It is also possible to set $\psi(X \rightarrow \alpha, (i, j, k))$ directly equal
 5587 to the probability $p(\alpha \mid X)$. In this case, \otimes is multiplication, and \bigoplus is addition.
 5588 While this may seem more intuitive than working with log probabilities, there is the
 5589 risk of underflow on long inputs.

5590 Regardless of how the scores are combined, the key point is the locality assumption:
 5591 the score for a derivation is the combination of the independent scores for each anchored

5592 production, and these scores do not depend on any other part of the derivation. For exam-
 5593 ple, if two non-terminals are siblings, the scores of productions from these non-terminals
 5594 are computed independently. This locality assumption is analogous to the first-order
 5595 Markov assumption in sequence labeling, where the score for transitions between tags
 5596 depends only on the previous tag and current tag, and not on the history. As with se-
 5597 quence labeling, this assumption makes it possible to find the optimal parse efficiently; its
 5598 linguistic limitations are discussed in § 10.5.

5599 10.4 Learning weighted context-free grammars

5600 Like sequence labeling, context-free parsing is a form of structure prediction. As a result,
 5601 WCFGs can be learned using the same set of algorithms: generative probabilistic models,
 5602 structured perceptron, maximum conditional likelihood, and maximum margin learning.
 5603 In all cases, learning requires a **treebank**, which is a dataset of sentences labeled with
 5604 context-free parses. Parsing research was catalyzed by the **Penn Treebank** (Marcus et al.,
 5605 1993), the first large-scale dataset of this type (see § 9.2.2). Phrase structure treebanks exist
 5606 for roughly two dozen other languages, with coverage mainly restricted to European and
 5607 East Asian languages, plus Arabic and Urdu.

5608 10.4.1 Probabilistic context-free grammars

Probabilistic context-free grammars are similar to hidden Markov models, in that they are generative models of text. In this case, the parameters of interest correspond to probabilities of productions, conditional on the left-hand side. As with hidden Markov models, these parameters can be estimated by relative frequency:

$$\psi(X \rightarrow \alpha) = \log p(X \rightarrow \alpha) \quad [10.19]$$

$$\hat{p}(X \rightarrow \alpha) = \frac{\text{count}(X \rightarrow \alpha)}{\text{count}(X)}. \quad [10.20]$$

5609 For example, the probability of the production $\text{NP} \rightarrow \text{DET NN}$ is the corpus count of
 5610 this production, divided by the count of the non-terminal NP. This estimator applies
 5611 to terminal productions as well: the probability of $\text{NN} \rightarrow \text{whale}$ is the count of how often
 5612 *whale* appears in the corpus as generated from an NN tag, divided by the total count of the
 5613 NN tag. Even with the largest treebanks — currently on the order of one million tokens
 5614 — it is difficult to accurately compute probabilities of even moderately rare events, such
 5615 as $\text{NN} \rightarrow \text{whale}$. Therefore, smoothing is critical for making PCFGs effective.

5616 **10.4.2 Feature-based parsing**

5617 The scores for each production can be computed as an inner product of weights and fea-
5618 tures,

$$\psi(X \rightarrow \alpha, (i, j, k)) = \boldsymbol{\theta} \cdot \mathbf{f}(X, \alpha, (i, j, k), \mathbf{w}), \quad [10.21]$$

5619 where the feature vector \mathbf{f} is a function of the left-hand side X , the right-hand side α , the
5620 anchor indices (i, j, k) , and the input \mathbf{w} .

5621 The basic feature $\mathbf{f}(X, \alpha, (i, j, k)) = \{(X, \alpha)\}$ encodes only the identity of the produc-
5622 tion itself. This gives rise to a discriminatively-trained model with the same expressive-
5623 ness as a PCFG. Features on anchored productions can include the words that border the
5624 span w_i, w_{j+1} , the word at the split point w_{k+1} , the presence of a verb or noun in the left
5625 child span $w_{i+1:k}$, and so on (Durrett and Klein, 2015). Scores on anchored productions
5626 can be incorporated into CKY parsing without any modification to the algorithm, because
5627 it is still possible to compute each element of the table $t[i, j, X]$ recursively from its imme-
5628 diate children.

5629 Other features can be obtained by grouping elements on either the left-hand or right-
5630 hand side: for example it can be particularly beneficial to compute additional features
5631 by clustering terminal symbols, with features corresponding to groups of words with
5632 similar syntactic properties. The clustering can be obtained from unlabeled datasets that
5633 are much larger than any treebank, improving coverage. Such methods are described in
5634 chapter 14.

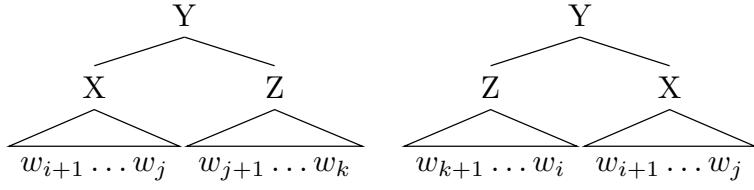
Feature-based parsing models can be estimated using the usual array of discriminative learning techniques. For example, a structure perceptron update can be computed as (Carreras et al., 2008),

$$\mathbf{f}(\tau, \mathbf{w}^{(i)}) = \sum_{(X \rightarrow \alpha, (i, j, k)) \in \tau} \mathbf{f}(X, \alpha, (i, j, k), \mathbf{w}^{(i)}) \quad [10.22]$$

$$\hat{\tau} = \underset{\tau \in \mathcal{T}(\mathbf{w})}{\operatorname{argmax}} \boldsymbol{\theta} \cdot \mathbf{f}(\tau, \mathbf{w}^{(i)}) \quad [10.23]$$

$$\boldsymbol{\theta} \leftarrow \mathbf{f}(\tau^{(i)}, \mathbf{w}^{(i)}) - \mathbf{f}(\hat{\tau}, \mathbf{w}^{(i)}). \quad [10.24]$$

5635 A margin-based objective can be optimized by selecting $\hat{\tau}$ through cost-augmented decod-
5636 ing (§ 2.4.2), enforcing a margin of $\Delta(\hat{\tau}, \tau)$ between the hypothesis and the reference parse,
5637 where Δ is a non-negative cost function, such as the Hamming loss (Stern et al., 2017). It
5638 is also possible to train feature-based parsing models by conditional log-likelihood, as
5639 described in the next section.

Figure 10.3: The two cases faced by the outside recurrence in the computation of $\beta(i, j, X)$

5640 **10.4.3 *Conditional random field parsing**

5641 The score of a derivation $\Psi(\tau)$ can be converted into a probability by normalizing over all
5642 possible derivations,

$$p(\tau \mid \mathbf{w}) = \frac{\exp \Psi(\tau)}{\sum_{\tau' \in \mathcal{T}(\mathbf{w})} \exp \Psi(\tau')}. \quad [10.25]$$

5643 Using this probability, a WCFG can be trained by maximizing the conditional log-likelihood
5644 of a labeled corpus.

5645 Just as in logistic regression and the conditional random field over sequences, the
5646 gradient of the conditional log-likelihood is the difference between the observed and ex-
5647 pected counts of each feature. The expectation $E_{\tau \mid \mathbf{w}}[\mathbf{f}(\tau, \mathbf{w}^{(i)}; \boldsymbol{\theta})]$ requires summing over
5648 all possible parses, and computing the marginal probabilities of anchored productions,
5649 $p(X \rightarrow \alpha, (i, j, k) \mid \mathbf{w})$. In CRF sequence labeling, marginal probabilities over tag bigrams
5650 are computed by the two-pass **forward-backward algorithm** (§ 7.5.3). The analogue for
5651 context-free grammars is the **inside-outside algorithm**, in which marginal probabilities
5652 are computed from terms generated by an upward and downward pass over the parsing
5653 chart:

- The upward pass is performed by the inside recurrence, which is described in § 10.3.2.
Each inside variable $\alpha(i, j, X)$ is the score of deriving $w_{i+1:j}$ from the non-terminal
 X . In a PCFG, this corresponds to the log-probability $\log p(w_{i+1:j} \mid X)$. This is
computed by the recurrence,

$$\alpha(i, j, X) \triangleq \log \sum_{(X \rightarrow Y \ Z)} \sum_{k=i+1}^j \exp (\psi(X \rightarrow Y \ Z, (i, j, k)) + \alpha(i, k, Y) + \alpha(k, j, Z)). \quad [10.26]$$

5654 The initial condition of this recurrence is $\alpha(m-1, m, X) = \psi(X \rightarrow w_m)$. The de-
5655 nominator $\sum_{\tau \in \mathcal{T}(\mathbf{w})} \exp \Psi(\tau)$ is equal to $\exp \alpha(0, M, S)$.

- The downward pass is performed by the **outside recurrence**, which recursively pop-
ulates the same table structure, starting at the root of the tree. Each outside variable

$\beta(i, j, X)$ is the score of having a phrase of type X covering the span $(i + 1 : j)$, joint with the exterior context $w_{1:i}$ and $w_{j+1:M}$. In a PCFG, this corresponds to the log probability $\log p((X, i + 1, j), w_{1:i}, w_{j+1:M})$. Each outside variable is computed by the recurrence,

$$\exp \beta(i, j, X) \triangleq \sum_{(Y \rightarrow X \mid Z)} \sum_{k=j+1}^M \exp [\psi(Y \rightarrow X \mid Z, (i, k, j)) + \alpha(j, k, Z) + \beta(i, k, Y)] \quad [10.27]$$

$$+ \sum_{(Y \rightarrow Z \mid X)} \sum_{k=0}^{i-1} \exp [\psi(Y \rightarrow Z \mid X, (k, i, j)) + \alpha(k, i, Z) + \beta(k, j, Y)]. \quad [10.28]$$

5656 The first line of Equation 10.28 is the score under the condition that X is a left child
 5657 of its parent, which spans $w_{i+1:k}$, with $k > j$; the second line is the score under the
 5658 condition that X is a right child of its parent Y , which spans $w_{k+1:j}$, with $k < i$.
 5659 The two cases are shown in Figure 10.3. In each case, we sum over all possible
 5660 productions with X on the right-hand side. The parent Y is bounded on one side
 5661 by either i or j , depending on whether X is a left or right child of Y ; we must sum
 5662 over all possible values for the other boundary. The initial conditions for the outside
 5663 recurrence are $\beta(0, M, S) = 0$ and $\beta(0, M, X \neq S) = -\infty$.

The marginal probability of a non-terminal X over span $w_{i+1:j}$ is written $p(X \rightsquigarrow w_{i+1:j} \mid \mathbf{w})$. This probability can be computed from the inside and outside scores,

$$p(X \rightsquigarrow w_{i+1:j} \mid \mathbf{w}) = \frac{p(X \rightsquigarrow w_{i+1:j}, \mathbf{w})}{p(\mathbf{w})} \quad [10.29]$$

$$= \frac{p(w_{i+1:j} \mid X) \times p(X, w_{1:i}, w_{j+1:M})}{p(\mathbf{w})} \quad [10.30]$$

$$= \frac{\exp(\alpha(i, j, X) + \beta(i, j, X))}{\exp \alpha(0, M, S)}. \quad [10.31]$$

5664 Marginal probabilities of individual productions can be computed similarly (see exercise
 5665 2). These marginal probabilities can be used for training a conditional random field parser,
 5666 and also for the task of unsupervised **grammar induction**, in which a PCFG is estimated
 5667 from a dataset of unlabeled text (Lari and Young, 1990; Pereira and Schabes, 1992).

5668 **10.4.4 Neural context-free grammars**

5669 Neural networks can be applied to parsing by representing each span with a dense
 5670 numerical vector (Socher et al., 2013; Durrett and Klein, 2015; Cross and Huang, 2016).⁴
 5671 For example, the anchor (i, j, k) and sentence w can be associated with a fixed-length
 5672 column vector,

$$\mathbf{v}_{(i,j,k)} = [\mathbf{u}_{w_{i-1}}; \mathbf{u}_{w_i}; \mathbf{u}_{w_{j-1}}; \mathbf{u}_{w_j}; \mathbf{u}_{w_{k-1}}; \mathbf{u}_{w_k}], \quad [10.32]$$

where \mathbf{u}_{w_i} is a word embedding associated with the word w_i . The vector $\mathbf{v}_{(i,j,k)}$ can then be passed through a feedforward neural network, and used to compute the score of the anchored production. For example, this score can be computed as a bilinear product (Durrett and Klein, 2015),

$$\tilde{\mathbf{v}}_{(i,j,k)} = \text{FeedForward}(\mathbf{v}_{(i,j,k)}) \quad [10.33]$$

$$\psi(X \rightarrow \alpha, (i, j, k)) = \tilde{\mathbf{v}}_{(i,j,k)}^\top \Theta \mathbf{f}(X \rightarrow \alpha), \quad [10.34]$$

5673 where $\mathbf{f}(X \rightarrow \alpha)$ is a vector of features of the production, and Θ is a parameter matrix.
 5674 The matrix Θ and the parameters of the feedforward network can be learned by
 5675 backpropagating from an objective such as the margin loss or the negative conditional
 5676 log-likelihood.

5677 **10.5 Grammar refinement**

5678 The locality assumptions underlying CFG parsing depend on the granularity of the non-
 5679 terminals. For the Penn Treebank non-terminals, there are several reasons to believe that
 5680 these assumptions are too strong (Johnson, 1998):

- 5681 • The context-free assumption is too strict: for example, the probability of the production
 5682 $\text{NP} \rightarrow \text{NP PP}$ is much higher (in the PTB) if the parent of the noun phrase is a
 5683 verb phrase (indicating that the NP is a direct object) than if the parent is a sentence
 5684 (indicating that the NP is the subject of the sentence).
- 5685 • The Penn Treebank non-terminals are too coarse: there are many kinds of noun
 5686 phrases and verb phrases, and accurate parsing sometimes requires knowing the
 5687 difference. As we have already seen, when faced with prepositional phrase at-
 5688 tachment ambiguity, a weighted CFG will either always choose NP attachment (if
 5689 $\psi(\text{NP} \rightarrow \text{NP PP}) > \psi(\text{VP} \rightarrow \text{VP PP})$), or it will always choose VP attachment. To
 5690 get more nuanced behavior, more fine-grained non-terminals are needed.
- 5691 • More generally, accurate parsing requires some amount of **semantics** — understand-
 5692 ing the meaning of the text to be parsed. Consider the example *cats scratch people*

⁴Earlier work on neural constituent parsing used transition-based parsing algorithms (§ 10.6.2) rather than CKY-style chart parsing (Henderson, 2004; Titov and Henderson, 2007).

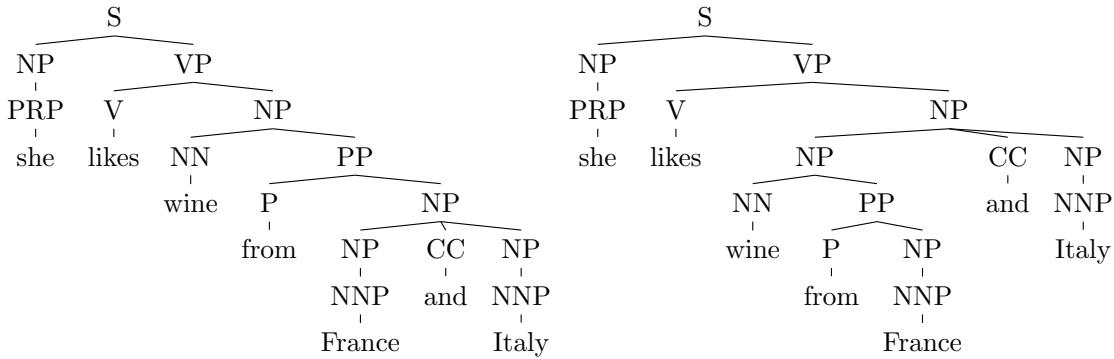


Figure 10.4: The left parse is preferable because of the conjunction of phrases headed by *France* and *Italy*, but these parses cannot be distinguished by a WCFG.

5693 *with claws*: knowledge of about cats, claws, and scratching is necessary to correctly
 5694 resolve the attachment ambiguity.

5695 An extreme example is shown in Figure 10.4. The analysis on the left is preferred
 5696 because of the conjunction of similar entities *France* and *Italy*. But given the non-terminals
 5697 shown in the analyses, there is no way to differentiate these two parses, since they include
 5698 exactly the same productions. What is needed seems to be more precise non-terminals.
 5699 One possibility would be to rethink the linguistics behind the Penn Treebank, and ask
 5700 the annotators to try again. But the original annotation effort took five years, and there
 5701 is a little appetite for another annotation effort of this scope. Researchers have therefore
 5702 turned to automated techniques.

5703 10.5.1 Parent annotations and other tree transformations

The key assumption underlying context-free parsing is that productions depend only on the identity of the non-terminal on the left-hand side, and not on its ancestors or neighbors. The validity of this assumption is an empirical question, and it depends on the non-terminals themselves: ideally, every noun phrase (and verb phrase, etc) would be distributionally identical, so the assumption would hold. But in the Penn Treebank, the observed probability of productions often depends on the parent of the left-hand side. For example, noun phrases are more likely to be modified by prepositional phrases when they are in the object position (e.g., *they amused the students from Georgia*) than in the subject position (e.g., *the students from Georgia amused them*). This means that the $\text{NP} \rightarrow \text{NP PP}$ production is more likely if the entire constituent is the child of a VP than if it is the child

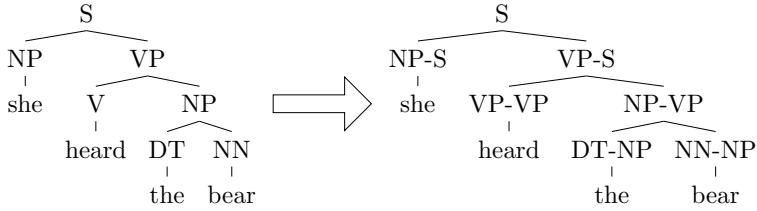


Figure 10.5: Parent annotation in a CFG derivation

of S. The observed statistics are (Johnson, 1998):

$$\Pr(\text{NP} \rightarrow \text{NP PP}) = 11\% \quad [10.35]$$

$$\Pr(\text{NP under } S \rightarrow \text{NP PP}) = 9\% \quad [10.36]$$

$$\Pr(\text{NP under } \text{VP} \rightarrow \text{NP PP}) = 23\%. \quad [10.37]$$

5704 This phenomenon can be captured by **parent annotation** (Johnson, 1998), in which each
 5705 non-terminal is augmented with the identity of its parent, as shown in Figure 10.5). This is
 5706 sometimes called **vertical Markovization**, since a Markov dependency is introduced be-
 5707 tween each node and its parent (Klein and Manning, 2003). It is analogous to moving from
 5708 a bigram to a trigram context in a hidden Markov model. In principle, parent annotation
 5709 squares the size of the set of non-terminals, which could make parsing considerably less
 5710 efficient. But in practice, the increase in the number of non-terminals that actually appear
 5711 in the data is relatively modest (Johnson, 1998).

5712 Parent annotation weakens the WCFG locality assumptions. This improves accuracy
 5713 by enabling the parser to make more fine-grained distinctions, which better capture real
 5714 linguistic phenomena. However, each production is more rare, and so careful smoothing
 5715 or regularization is required to control the variance over production scores.

5716 10.5.2 Lexicalized context-free grammars

5717 The examples in § 10.2.2 demonstrate the importance of individual words in resolving
 5718 parsing ambiguity: the preposition *on* is more likely to attach to *met*, while the preposition
 5719 *of* is more likely to attachment to *President*. But of all word pairs, which are relevant to
 5720 attachment decisions? Consider the following variants on the original examples:

- 5721 (10.3) a. We met the President of Mexico.
 5722 b. We met the first female President of Mexico.
 5723 c. They had supposedly met the President on Monday.

5724 The underlined words are the **head words** of their respective phrases: *met* heads the verb
 5725 phrase, and *President* heads the direct object noun phrase. These heads provide useful

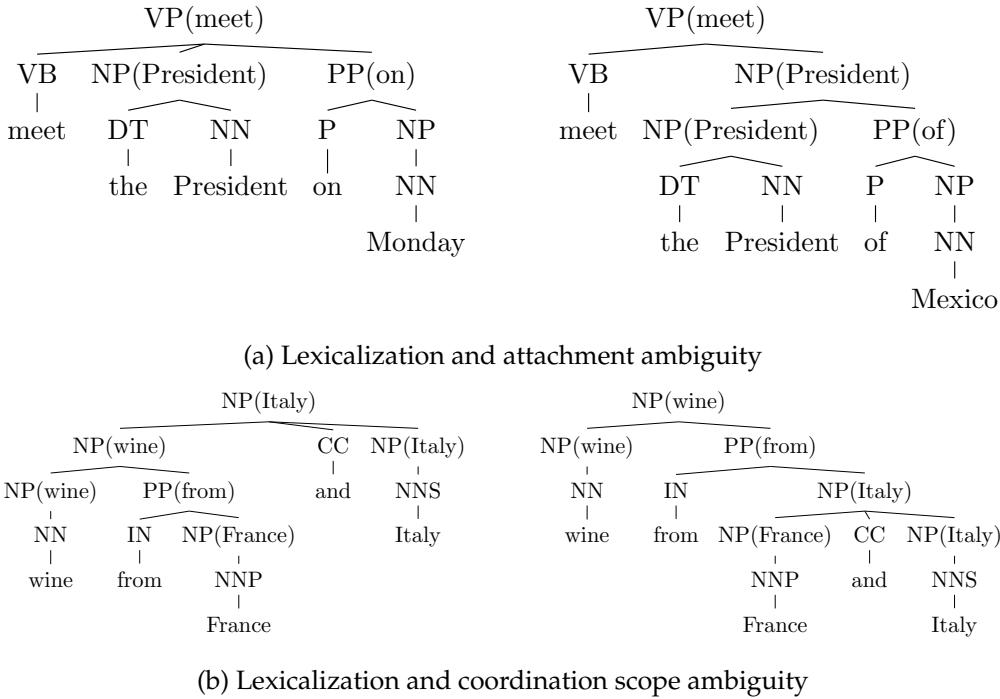


Figure 10.6: Examples of lexicalization

semantic information. But they break the context-free assumption, which states that the score for a production depends only on the parent and its immediate children, and not the substructure under each child.

The incorporation of head words into context-free parsing is known as **lexicalization**, and is implemented in rules of the form,

$$\text{NP}(President) \rightarrow \text{NP}(President) \text{ PP}(of) \quad [10.38]$$

$$\text{NP}(President) \rightarrow \text{NP}(President) \text{ PP}(on). \quad [10.39]$$

Lexicalization was a major step towards accurate PCFG parsing in the 1990s and early 2000s. It requires solving three problems: identifying the heads of all constituents in a treebank; parsing efficiently while keeping track of the heads; and estimating the scores for lexicalized productions.

Non-terminal	Direction	Priority
S	right	VP SBAR ADJP UCP NP
VP	left	VBD VBN MD VBZ TO VB VP VBG VBP ADJP NP
NP	right	N* EX \$ CD QP PRP ...
PP	left	IN TO FW

Table 10.3: A fragment of head percolation rules for English (Magerman, 1995; Collins, 1997)

5733 Identifying head words

5734 The head of a constituent is the word that is the most useful for determining how that
 5735 constituent is integrated into the rest of the sentence.⁵ The head word of a constituent is
 5736 determined recursively: for any non-terminal production, the head of the left-hand side
 5737 must be the head of one of the children. The head is typically selected according to a set of
 5738 deterministic rules, sometimes called **head percolation rules**. In many cases, these rules
 5739 are straightforward: the head of a noun phrase in a $NP \rightarrow DET\ NN$ production is the head
 5740 of the noun; the head of a sentence in a $S \rightarrow NP\ VP$ production is the head of the verb
 5741 phrase.

5742 Table 10.3 shows a fragment of the head percolation rules used in many English pars-
 5743 ing systems. The meaning of the first rule is that to find the head of an S constituent, first
 5744 look for the rightmost VP child; if you don't find one, then look for the rightmost SBAR
 5745 child, and so on down the list. Verb phrases are headed by left verbs (the head of *can plan*
 5746 *on walking* is *planned*, since the modal verb *can* is tagged MD); noun phrases are headed by
 5747 the rightmost noun-like non-terminal (so the head of *the red cat* is *cat*),⁶ and prepositional
 5748 phrases are headed by the preposition (the head of *at Georgia Tech* is *at*). Some of these
 5749 rules are somewhat arbitrary — there's no particular reason why the head of *cats and dogs*
 5750 should be *dogs* — but the point here is just to get some lexical information that can support
 5751 parsing, not to make deep claims about syntax. Figure 10.6 shows the application of these
 5752 rules to two of the running examples.

5753 Parsing lexicalized context-free grammars

5754 A naïve application of lexicalization would simply increase the set of non-terminals by
 5755 taking the cross-product with the set of terminal symbols, so that the non-terminals now

⁵This is a pragmatic definition, befitting our goal of using head words to improve parsing; for a more formal definition, see (Bender, 2013, chapter 7).

⁶The noun phrase non-terminal is sometimes treated as a special case. Collins (1997) uses a heuristic that looks for the rightmost child which is a noun-like part-of-speech (e.g., NN, NNP), a possessive marker, or a superlative adjective (e.g., *the greatest*). If no such child is found, the heuristic then looks for the *leftmost* NP. If there is no child with tag NP, the heuristic then applies another priority list, this time from right to left.

5756 include symbols like $\text{NP}(\text{President})$ and $\text{VP}(\text{meet})$. Under this approach, the CKY parsing
 5757 algorithm could be applied directly to the lexicalized production rules. However, the
 5758 complexity would be cubic in the size of the vocabulary of terminal symbols, which would
 5759 clearly be intractable.

Another approach is to augment the CKY table with an additional index, keeping track of the head of each constituent. The cell $t[i, j, h, X]$ stores the score of the best derivation in which non-terminal X spans $w_{i+1:j}$ with head word h , where $i < h \leq j$. To compute such a table recursively, we must consider the possibility that each phrase gets its head from either its left or right child. The scores of the best derivations in which the head comes from the left and right child are denoted t_ℓ and t_r respectively, leading to the following recurrence:

$$t_\ell[i, j, h, X] = \max_{(X \rightarrow YZ)} \max_{k > h} \max_{k < h' \leq j} t[i, k, h, Y] + t[k, j, h', Z] + \psi(X(h) \rightarrow Y(h)Z(h')) \quad [10.40]$$

$$t_r[i, j, h, X] = \max_{(X \rightarrow YZ)} \max_{k < h} \max_{i < h' \leq k} t[i, k, h', Y] + t[k, j, h, Z] + (\psi(X(h) \rightarrow Y(h')Z(h))) \quad [10.41]$$

$$t[i, j, h, X] = \max(t_\ell[i, j, h, X], t_r[i, j, h, X]). \quad [10.42]$$

5760 To compute t_ℓ , we maximize over all split points $k > h$, since the head word must be in
 5761 the left child. We then maximize again over possible head words h' for the right child. An
 5762 analogous computation is performed for t_r . The size of the table is now $\mathcal{O}(M^3N)$, where
 5763 M is the length of the input and N is the number of non-terminals. Furthermore, each
 5764 cell is computed by performing $\mathcal{O}(M^2)$ operations, since we maximize over both the split
 5765 point k and the head h' . The time complexity of the algorithm is therefore $\mathcal{O}(RM^5N)$,
 5766 where R is the number of rules in the grammar. Fortunately, more efficient solutions are
 5767 possible. In general, the complexity of parsing can be reduced to $\mathcal{O}(M^4)$ in the length of
 5768 the input; for a broad class of lexicalized CFGs, the complexity can be made cubic in the
 5769 length of the input, just as in unlexicalized CFGs (Eisner, 2000).

5770 Estimating lexicalized context-free grammars

5771 The final problem for lexicalized parsing is how to estimate weights for lexicalized pro-
 5772 ductions $X(i) \rightarrow Y(j) Z(k)$. These productions are said to be bilexical, because they
 5773 involve scores over pairs of words: in the example *meet the President of Mexico*, we hope
 5774 to choose the correct attachment point by modeling the bilexical affinities of (*meet, of*) and
 5775 (*President, of*). The number of such word pairs is quadratic in the size of the vocabulary,
 5776 making it difficult to estimate the weights of lexicalized production rules directly from
 5777 data. This is especially true for probabilistic context-free grammars, in which the weights
 5778 are obtained from smoothed relative frequency. In a treebank with a million tokens, a

5779 vanishingly small fraction of the possible lexicalized productions will be observed more
 5780 than once.⁷ The Charniak (1997) and Collins (1997) parsers therefore focus on approximating
 5781 the probabilities of lexicalized productions, using various smoothing techniques
 5782 and independence assumptions.

In discriminatively-trained weighted context-free grammars, the scores for each production can be computed from a set of features, which can be made progressively more fine-grained (Finkel et al., 2008). For example, the score of the lexicalized production $\text{NP}(\text{President}) \rightarrow \text{NP}(\text{President}) \text{ PP}(of)$ can be computed from the following features:

$$\begin{aligned} f(\text{NP}(\text{President}) \rightarrow \text{NP}(\text{President}) \text{ PP}(of)) = & \{\text{NP}(*) \rightarrow \text{NP}(*) \text{ PP}(*), \\ & \text{NP}(\text{President}) \rightarrow \text{NP}(\text{President}) \text{ PP}(*), \\ & \text{NP}(*) \rightarrow \text{NP}(*) \text{ PP}(of), \\ & \text{NP}(\text{President}) \rightarrow \text{NP}(\text{President}) \text{ PP}(of)\} \end{aligned}$$

5783 The first feature scores the unlexicalized production $\text{NP} \rightarrow \text{NP PP}$; the next two features
 5784 lexicalize only one element of the production, thereby scoring the appropriateness of NP
 5785 attachment for the individual words *President* and *of*; the final feature scores the specific
 5786 blexical affinity of *President* and *of*. For blexical pairs that are encountered frequently in
 5787 the treebank, this blexical feature can play an important role in parsing; for pairs that are
 5788 absent or rare, regularization will drive its weight to zero, forcing the parser to rely on the
 5789 more coarse-grained features.

5790 In chapter 14, we will encounter techniques for clustering words based on their **distribu-**
 5791 **tional** properties — the contexts in which they appear. Such a clustering would group
 5792 rare and common words, such as *whale*, *shark*, *beluga*, *Leviathan*. Word clusters can be used
 5793 as features in discriminative lexicalized parsing, striking a middle ground between full
 5794 lexicalization and non-terminals (Finkel et al., 2008). In this way, labeled examples con-
 5795 taining relatively common words like *whale* can help to improve parsing for rare words
 5796 like *beluga*, as long as those two words are clustered together.

5797 10.5.3 *Refinement grammars

5798 Lexicalization improves on context-free parsing by adding detailed information in the
 5799 form of lexical heads. However, estimating the scores of lexicalized productions is dif-
 5800 ficult. Klein and Manning (2003) argue that the right level of linguistic detail is some-
 5801 where between treebank categories and individual words. Some parts-of-speech and non-
 5802 terminals are truly substitutable: for example, *cat*/N and *dog*/N. But others are not: for
 5803 example, the preposition *of* exclusively attaches to nouns, while the preposition *as* is more

⁷The real situation is even more difficult, because non-binary context-free grammars can involve trilexical or higher-order dependencies, between the head of the constituent and multiple of its children (Carreras et al., 2008).

likely to modify verb phrases. Klein and Manning (2003) obtained a 2% improvement in *F*-MEASURE on a parent-annotated PCFG parser by making a single change: splitting the preposition category into six subtypes. They propose a series of linguistically-motivated refinements to the Penn Treebank annotations, which in total yielded a 40% error reduction.

Non-terminal refinement process can be automated by treating the refined categories as **latent variables**. For example, we might split the noun phrase non-terminal into NP1, NP2, NP3, . . . , without defining in advance what each refined non-terminal corresponds to. This can be treated as partially supervised learning, similar to the multi-component document classification model described in § 5.2.3. A latent variable PCFG can be estimated by expectation maximization (Matsuzaki et al., 2005):⁸

- In the E-step, estimate a marginal distribution q over the refinement type of each non-terminal in each derivation. These marginals are constrained by the original annotation: an NP can be reannotated as NP4, but not as VP3. Marginal probabilities over refined productions can be computed from the **inside-outside algorithm**, as described in § 10.4.3, where the E-step enforces the constraints imposed by the original annotations.
- In the M-step, recompute the parameters of the grammar, by summing over the probabilities of anchored productions that were computed in the E-step:

$$E[\text{count}(X \rightarrow Y Z)] = \sum_{i=0}^M \sum_{j=i}^M \sum_{k=i}^j p(X \rightarrow Y Z, (i, j, k) \mid \mathbf{w}). \quad [10.43]$$

As usual, this process can be iterated to convergence. To determine the number of refinement types for each tag, Petrov et al. (2006) apply a split-merge heuristic; Liang et al. (2007) and Finkel et al. (2007) apply **Bayesian nonparametrics** (Cohen, 2016).

Some examples of refined non-terminals are shown in Table 10.4. The proper nouns differentiate months, first names, middle initials, last names, first names of places, and second names of places; each of these will tend to appear in different parts of grammatical productions. The personal pronouns differentiate grammatical role, with PRP-0 appearing in subject position at the beginning of the sentence (note the capitalization), PRP-1 appearing in subject position but not at the beginning of the sentence, and PRP-2 appearing in object position.

10.6 Beyond context-free parsing

In the context-free setting, the score for a parse is a combination of the scores of individual productions. As we have seen, these models can be improved by using finer-grained non-

⁸Spectral learning, described in § 5.5.2, has also been applied to refinement grammars (Cohen et al., 2014).

Proper nouns			
NNP-14	<i>Oct.</i>	<i>Nov.</i>	<i>Sept.</i>
NNP-12	<i>John</i>	<i>Robert</i>	<i>James</i>
NNP-2	<i>J.</i>	<i>E.</i>	<i>L.</i>
NNP-1	<i>Bush</i>	<i>Noriega</i>	<i>Peters</i>
NNP-15	<i>New</i>	<i>San</i>	<i>Wall</i>
NNP-3	<i>York</i>	<i>Francisco</i>	<i>Street</i>
Personal Pronouns			
PRP-0	<i>It</i>	<i>He</i>	<i>I</i>
PRP-1	<i>it</i>	<i>he</i>	<i>they</i>
PRP-2	<i>it</i>	<i>them</i>	<i>him</i>

Table 10.4: Examples of automatically refined non-terminals and some of the words that they generate (Petrov et al., 2006).

5836 terminals, via parent-annotation, lexicalization, and automated refinement. However, the
 5837 inherent limitations to the expressiveness of context-free parsing motivate the consider-
 5838 ation of other search strategies. These strategies abandon the optimality guaranteed by
 5839 bottom-up parsing, in exchange for the freedom to consider arbitrary properties of the
 5840 proposed parses.

5841 10.6.1 Reranking

5842 A simple way to relax the restrictions of context-free parsing is to perform a two-stage pro-
 5843 cess, in which a context-free parser generates a k -best list of candidates, and a **reranker**
 5844 then selects the best parse from this list (Charniak and Johnson, 2005; Collins and Koo,
 5845 2005). The reranker can be trained from an objective that is similar to multi-class classi-
 5846 fication: the goal is to learn weights that assign a high score to the reference parse, or to
 5847 the parse on the k -best list that has the lowest error. In either case, the reranker need only
 5848 evaluate the K best parses, and so no context-free assumptions are necessary. This opens
 5849 the door to more expressive scoring functions:

- 5850 • It is possible to incorporate arbitrary non-local features, such as the structural par-
 5851 allelism and right-branching orientation of the parse (Charniak and Johnson, 2005).
- 5852 • Reranking enables the use of **recursive neural networks**, in which each constituent
 5853 span $w_{i+1:j}$ receives a vector $u_{i,j}$ which is computed from the vector representa-
 5854 tions of its children, using a composition function that is linked to the production

5855 rule (Socher et al., 2013), e.g.,

$$\mathbf{u}_{i,j} = f\left(\Theta_{X \rightarrow Y \mid Z} \begin{bmatrix} \mathbf{u}_{i,k} \\ \mathbf{u}_{k,j} \end{bmatrix}\right) \quad [10.44]$$

5856 The overall score of the parse can then be computed from the final vector, $\Psi(\tau) =$
 5857 $\theta \mathbf{u}_{0,M}$.

5858 Reranking can yield substantial improvements in accuracy. The main limitation is that it
 5859 can only find the best parse among the K -best offered by the generator, so it is inherently
 5860 limited by the ability of the bottom-up parser to find high-quality candidates.

5861 10.6.2 Transition-based parsing

5862 Structure prediction can be viewed as a form of search. An alternative to bottom-up pars-
 5863 ing is to read the input from left-to-right, gradually building up a parse structure through
 5864 a series of **transitions**. Transition-based parsing is described in more detail in the next
 5865 chapter, in the context of dependency parsing. However, it can also be applied to CFG
 5866 parsing, as briefly described here.

5867 For any context-free grammar, there is an equivalent **pushdown automaton**, a model
 5868 of computation that accepts exactly those strings that can be derived from the grammar.
 5869 This computational model consumes the input from left to right, while pushing and pop-
 5870 ping elements on a stack. This architecture provides a natural transition-based parsing
 5871 framework for context-free grammars, known as **shift-reduce parsing**.

5872 Shift-reduce parsing is a type of transition-based parsing, in which the parser can take
 5873 the following actions:

- 5874 • *shift* the next terminal symbol onto the stack;
- 5875 • *unary-reduce* the top item on the stack, using a unary production rule in the gram-
 5876 mar;
- 5877 • *binary-reduce* the top two items onto the stack, using a binary production rule in the
 5878 grammar.

5879 The set of available actions is constrained by the situation: the parser can only shift if
 5880 there are remaining terminal symbols in the input, and it can only reduce if an applicable
 5881 production rule exists in the grammar. If the parser arrives at a state where the input
 5882 has been completely consumed, and the stack contains only the element S, then the input
 5883 is accepted. If the parser arrives at a non-accepting state where there are no possible
 5884 actions, the input is rejected. A parse error occurs if there is some action sequence that
 5885 would accept an input, but the parser does not find it.

5886 **Example** Consider the input *we eat sushi* and the grammar in Table 10.1. The input can
 5887 be parsed through the following sequence of actions:

- 5888 1. **Shift** the first token *we* onto the stack.
- 5889 2. **Reduce** the top item on the stack to NP, using the production $NP \rightarrow we$.
- 5890 3. **Shift** the next token *eat* onto the stack, and **reduce** it to V with the production $V \rightarrow$
 5891 *eat*.
- 5892 4. **Shift** the final token *sushi* onto the stack, and **reduce** it to NP. The input has been
 5893 completely consumed, and the stack contains [NP, V, NP].
- 5894 5. **Reduce** the top two items using the production $VP \rightarrow V NP$. The stack now con-
 5895 tains [VP, NP].
- 5896 6. **Reduce** the top two items using the production $S \rightarrow NP VP$. The stack now contains
 5897 [S]. Since the input is empty, this is an accepting state.

5898 One thing to notice from this example is that the number of shift actions is equal to the
 5899 length of the input. The number of reduce actions is equal to the number of non-terminals
 5900 in the analysis, which grows linearly in the length of the input. Thus, the overall time
 5901 complexity of shift-reduce parsing is linear in the length of the input (assuming the com-
 5902 plexity of each individual classification decision is constant in the length of the input).
 5903 This is far better than the cubic time complexity required by CKY parsing.

5904 **Transition-based parsing as inference** In general, it is not possible to guarantee that
 5905 a transition-based parser will find the optimal parse, $\text{argmax}_\tau \Psi(\tau; \mathbf{w})$, even under the
 5906 usual CFG independence assumptions. We could assign a score to each anchored parsing
 5907 action in each context, with $\psi(a, c)$ indicating the score of performing action a in context c .
 5908 One might imagine that transition-based parsing could efficiently find the derivation that
 5909 maximizes the sum of such scores. But this too would require backtracking and searching
 5910 over an exponentially large number of possible action sequences: if a bad decision is
 5911 made at the beginning of the derivation, then it may be impossible to recover the optimal
 5912 action sequence without backtracking to that early mistake. This is known as a **search**
 5913 **error**. Transition-based parsers can incorporate arbitrary features, without the restrictive
 5914 independence assumptions required by chart parsing; search errors are the price that must
 5915 be paid for this flexibility.

5916 **Learning transition-based parsing** Transition-based parsing can be combined with ma-
 5917 chine learning by training a classifier to select the correct action in each situation. This
 5918 classifier is free to choose any feature of the input, the state of the parser, and the parse
 5919 history. However, there is no optimality guarantee: the parser may choose a suboptimal
 5920 parse, due to a mistake at the beginning of the analysis. Nonetheless, some of the strongest

5921 CFG parsers are based on the shift-reduce architecture, rather than CKY. A recent genera-
 5922 tion of models links shift-reduce parsing with recurrent neural networks, updating a
 5923 hidden state vector while consuming the input (e.g., Cross and Huang, 2016; Dyer et al.,
 5924 2016). Learning algorithms for transition-based parsing are discussed in more detail in
 5925 § 11.3.

5926 **Exercises**

5927 1. Design a grammar that handles English subject-verb agreement. Specifically, your
 5928 grammar should handle the examples below correctly:

5929 (10.4) a. She sings.

5930 b. We sing.

5931 (10.5) a. *She sing.

5932 b. *We sings.

5933 2. Extend your grammar from the previous problem to include the auxiliary verb *can*,
 5934 so that the following cases are handled:

5935 (10.6) a. She can sing.

5936 b. We can sing.

5937 (10.7) a. *She can sings.

5938 b. *We can sings.

5939 3. French requires subjects and verbs to agree in person and number, and it requires
 5940 determiners and nouns to agree in gender and number. Verbs and their objects need
 5941 not agree. Assuming that French has two genders (feminine and masculine), three
 5942 persons (first [*me*], second [*you*], third [*her*]), and two numbers (singular and plural),
 5943 how many productions are required to extend the following simple grammar to
 5944 handle agreement?

S	\rightarrow	NP VP
VP	\rightarrow	V V NP V NP NP
NP	\rightarrow	DET NN

5946 4. Consider the grammar:

5947	S → NP VP VP → V NP NP → JJ NP NP → <i>fish</i> (the animal) V → <i>fish</i> (the action of fishing) JJ → <i>fish</i> (a modifier, as in <i>fish sauce</i> or <i>fish stew</i>)
------	---

5948 Apply the CKY algorithm and identify all possible parses for the sentence *fish fish fish*.
 5949

- 5950 5. Choose one of the possible parses for the previous problem, and show how it can be
 5951 derived by a series of shift-reduce actions.
- 5952 6. To handle VP coordination, a grammar includes the production $VP \rightarrow VP\ CC\ VP$.
 5953 To handle adverbs, it also includes the production $VP \rightarrow VP\ ADV$. Assume all verbs
 5954 are generated from a sequence of unary productions, e.g., $VP \rightarrow V \rightarrow eat$.
- 5955 a) Show how to binarize the production $VP \rightarrow VP\ CC\ VP$.
- 5956 b) Use your binarized grammar to parse the sentence *They eat and drink together*,
 5957 treating *together* as an adverb.
- 5958 c) Prove that a weighted CFG cannot distinguish the two possible derivations of
 5959 this sentence. Your explanation should focus on the productions in the original,
 5960 non-binary grammar.
- 5961 d) Explain what condition must hold for a parent-annotated WCFG to prefer the
 5962 derivation in which *together* modifies the coordination *eat and drink*.

7. Consider the following PCFG:

$$p(X \rightarrow X\ X) = \frac{1}{2} \quad [10.45]$$

$$p(X \rightarrow Y) = \frac{1}{2} \quad [10.46]$$

$$p(Y \rightarrow \sigma) = \frac{1}{|\Sigma|}, \forall \sigma \in \Sigma \quad [10.47]$$

- 5963 a) Compute the probability $p(\hat{\tau})$ of the maximum probability parse for a string
 5964 $w \in \Sigma^M$.
 5965 b) Compute the conditional probability $p(\hat{\tau} | w)$.
- 5966 8. Context-free grammars can be used to parse the internal structure of words. Us-
 5967 ing the weighted CKY algorithm and the following weighted context-free grammar,
 5968 identify the best parse for the sequence of morphological segments *in+flame+able*.

	S	→	V	0
	S	→	N	0
	S	→	J	0
	V	→	VPref N	-1
5969	J	→	N JSuff	1
	J	→	V JSuff	0
	J	→	NegPref J	1
	VPref	→	in+	2
	NegPref	→	in+	1
	N	→	flame	0
	JSuff	→	+able	0

- 5970 9. Use the inside and outside scores to compute the marginal probability $p(X_{i+1:j} \rightarrow Y_{i+1:k} Z_{k+1:j} \mid \mathbf{w})$,
 5971 indicating that Y spans $\mathbf{w}_{i+1:k}$, Z spans $\mathbf{w}_{k+1:j}$, and X is the parent of Y and Z , span-
 5972 ning $\mathbf{w}_{i+1:j}$.
- 5973 10. Suppose that the potentials $\Psi(X \rightarrow \alpha)$ are log-probabilities, so that $\sum_{\alpha} \exp \Psi(X \rightarrow \alpha) = 1$
 5974 for all X . Verify that the semiring inside recurrence from Equation 10.26 generates
 5975 the log-probability $\log p(\mathbf{w}) = \log \sum_{\tau: \text{yield}(\tau) = \mathbf{w}} p(\tau)$.

5976 Chapter 11

5977 Dependency parsing

5978 The previous chapter discussed algorithms for analyzing sentences in terms of nested con-
5979 stituents, such as noun phrases and verb phrases. However, many of the key sources of
5980 ambiguity in phrase-structure analysis relate to questions of **attachment**: where to attach a
5981 prepositional phrase or complement clause, how to scope a coordinating conjunction, and
5982 so on. These attachment decisions can be represented with a more lightweight structure:
5983 a directed graph over the words in the sentence, known as a **dependency parse**. Syntac-
5984 tic annotation has shifted its focus to such dependency structures: at the time of this
5985 writing, the **Universal Dependencies** project offers more than 100 dependency treebanks
5986 for more than 60 languages.¹ This chapter will describe the linguistic ideas underlying
5987 dependency grammar, and then discuss exact and transition-based parsing algorithms.
5988 The chapter will also discuss recent research on **learning to search** in transition-based
5989 structure prediction.

5990 11.1 Dependency grammar

5991 While **dependency grammar** has a rich history of its own (Tesnière, 1966; Kübler et al.,
5992 2009), it can be motivated by extension from the lexicalized context-free grammars that
5993 we encountered in previous chapter (§ 10.5.2). Recall that lexicalization augments each
5994 non-terminal with a **head word**. The head of a constituent is identified recursively, using
5995 a set of **head rules**, as shown in Table 10.3. An example of a lexicalized context-free parse
5996 is shown in Figure 11.1a. In this sentence, the head of the S constituent is the main verb,
5997 *scratch*; this non-terminal then produces the noun phrase *the cats*, whose head word is
5998 *cats*, and from which we finally derive the word *the*. Thus, the word *scratch* occupies the
5999 central position for the sentence, with the word *cats* playing a supporting role. In turn, *cats*

¹universaldependencies.org

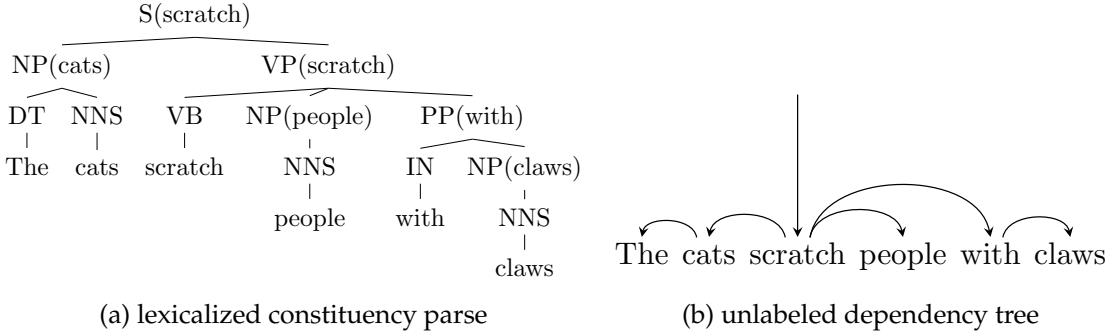


Figure 11.1: Dependency grammar is closely linked to lexicalized context free grammars: each lexical head has a dependency path to every other word in the constituent. (This example is based on the lexicalization rules from § 10.5.2, which make the preposition the head of a prepositional phrase. In the more contemporary Universal Dependencies annotations, the head of *with claws* would be *claws*, so there would be an edge *scratch* → *claws*.)

6000 occupies the central position for the noun phrase, with the word *the* playing a supporting
6001 role.

6002 The relationships between words in a sentence can be formalized in a directed graph,
6003 based on the lexicalized phrase-structure parse: create an edge (i, j) iff word i is the head
6004 of a phrase whose child is a phrase headed by word j . Thus, in our example, we would
6005 have *scratch* → *cats* and *cats* → *the*. We would not have the edge *scratch* → *the*, because
6006 although $S(\text{scratch})$ dominates $\text{DET}(\text{the})$ in the phrase-structure parse tree, it is not its im-
6007 mediate parent. These edges describe **syntactic dependencies**, a bilexical relationship
6008 between a **head** and a **dependent**, which is at the heart of dependency grammar.

6009 Continuing to build out this **dependency graph**, we will eventually reach every word
6010 in the sentence, as shown in Figure 11.1b. In this graph — and in all graphs constructed
6011 in this way — every word has exactly one incoming edge, except for the root word, which
6012 is indicated by a special incoming arrow from above. Furthermore, the graph is *weakly*
6013 *connected*: if the directed edges were replaced with undirected edges, there would be a
6014 path between all pairs of nodes. From these properties, it can be shown that there are no
6015 cycles in the graph (or else at least one node would have to have more than one incoming
6016 edge), and therefore, the graph is a tree. Because the graph includes all vertices, it is a
6017 **spanning tree**.

6018 11.1.1 Heads and dependents

6019 A dependency edge implies an asymmetric syntactic relationship between the head and
6020 dependent words, sometimes called **modifiers**. For a pair like *the cats* or *cats scratch*, how

6021 do we decide which is the head? Here are some possible criteria:

- 6022 • The head sets the syntactic category of the construction: for example, nouns are the
6023 heads of noun phrases, and verbs are the heads of verb phrases.
- 6024 • The modifier may be optional while the head is mandatory: for example, in the
6025 sentence *cats scratch people with claws*, the subtrees *cats scratch* and *cats scratch people*
6026 are grammatical sentences, but *with claws* is not.
- 6027 • The head determines the morphological form of the modifier: for example, in lan-
6028 guages that require gender agreement, the gender of the noun determines the gen-
6029 der of the adjectives and determiners.
- 6030 • Edges should first connect content words, and then connect function words.

6031 These guidelines are not universally accepted, and they sometimes conflict. The Uni-
6032 versal Dependencies (UD) project has attempted to identify a set of principles that can be
6033 applied to dozens of different languages (Nivre et al., 2016).² These guidelines are based
6034 on the universal part-of-speech tags from chapter 8. They differ somewhat from the head
6035 rules described in § 10.5.2: for example, on the principle that dependencies should relate
6036 content words, the prepositional phrase *with claws* would be headed by *claws*, resulting in
6037 an edge *scratch → claws*, and another edge *claws → with*.

6038 One objection to dependency grammar is that not all syntactic relations are asymmet-
6039 ric. One such relation is coordination (Popel et al., 2013): in the sentence, *Abigail and Max*
6040 *like kimchi* (Figure 11.2), which word is the head of the coordinated noun phrase *Abigail*
6041 and *Max*? Choosing either *Abigail* or *Max* seems arbitrary; fairness argues for making *and*
6042 the head, but this seems like the least important word in the noun phrase, and selecting
6043 it would violate the principle of linking content words first. The Universal Dependencies
6044 annotation system arbitrarily chooses the left-most item as the head — in this case, *Abigail*
6045 — and includes edges from this head to both *Max* and the coordinating conjunction *and*.
6046 These edges are distinguished by the labels CONJ (for the thing begin conjoined) and CC
6047 (for the coordinating conjunction). The labeling system is discussed next.

6048 11.1.2 Labeled dependencies

6049 Edges may be **labeled** to indicate the nature of the syntactic relation that holds between
6050 the two elements. For example, in Figure 11.2, the label NSUBJ on the edge from *like* to
6051 *Abigail* indicates that the subtree headed by *Abigail* is the noun subject of the verb *like*;
6052 similarly, the label OBJ on the edge from *like* to *kimchi* indicates that the subtree headed by

²The latest and most specific guidelines are available at universaldependencies.org/guidelines.html

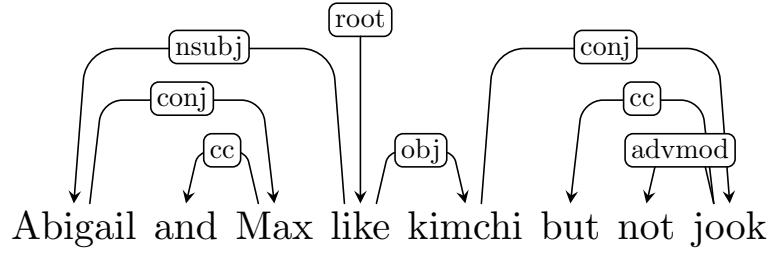


Figure 11.2: In the Universal Dependencies annotation system, the left-most item of a coordination is the head.

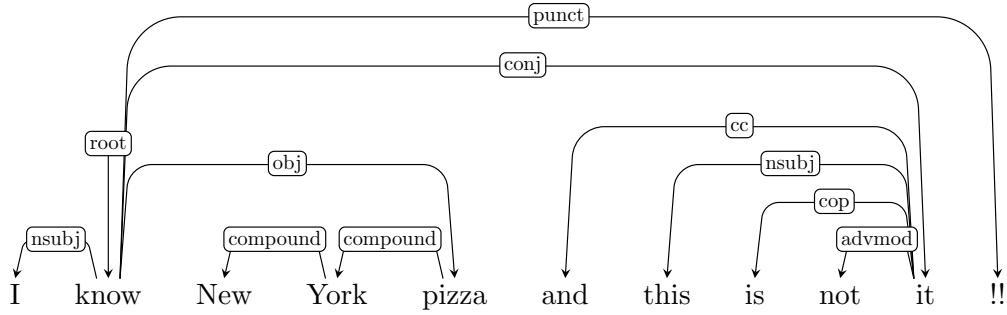


Figure 11.3: A labeled dependency parse from the English UD Treebank (reviews-361348-0006)

6053 *kimchi* is the object.³ The negation *not* is treated as an adverbial modifier (ADVMOD) on
6054 the noun *jook*.

6055 A slightly more complex example is shown in Figure 11.3. The multiword expression
6056 *New York pizza* is treated as a “flat” unit of text, with the elements linked by the COM-
6057 POUND relation. The sentence includes two clauses that are conjoined in the same way
6058 that noun phrases are conjoined in Figure 11.2. The second clause contains a **copula** verb
6059 (see § 8.1.1). For such clauses, we treat the “object” of the verb as the root — in this case,
6060 *it* — and label the verb as a dependent, with the COP relation. This example also shows
6061 how punctuation are treated, with label PUNCT.

6062 11.1.3 Dependency subtrees and constituents

6063 Dependency trees hide information that would be present in a CFG parse. Often what
6064 is hidden is in fact irrelevant: for example, Figure 11.4 shows three different ways of

³Earlier work distinguished direct and indirect objects (De Marneffe and Manning, 2008), but this has been dropped in version 2.0 of the Universal Dependencies annotation system.

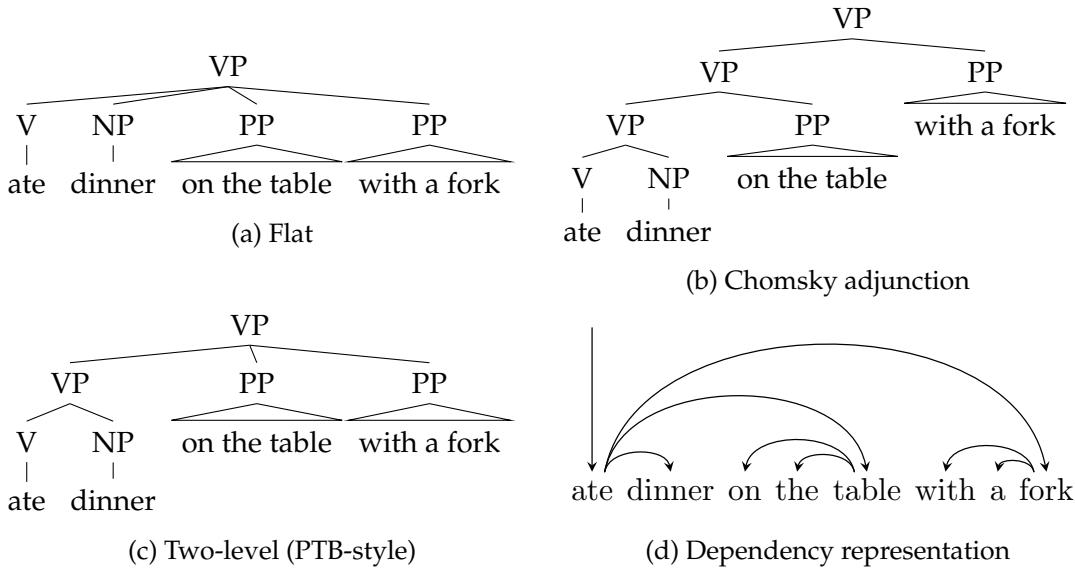


Figure 11.4: The three different CFG analyses of this verb phrase all correspond to a single dependency structure.

representing prepositional phrase adjuncts to the verb *ate*. Because there is apparently no meaningful difference between these analyses, the Penn Treebank decides by convention to use the two-level representation (see Johnson, 1998, for a discussion). As shown in Figure 11.4d, these three cases all look the same in a dependency parse.

But dependency grammar imposes its own set of annotation decisions, such as the identification of the head of a coordination (§ 11.1.1); without lexicalization, context-free grammar does not require either element in a coordination to be privileged in this way. Dependency parses can be disappointingly flat: for example, in the sentence *Yesterday, Abigail was reluctantly giving Max kimchi*, the root *giving* is the head of every dependency! The constituent parse arguably offers a more useful structural analysis for such cases.

Projectivity Thus far, we have defined dependency trees as spanning trees over a graph in which each word is a vertex. As we have seen, one way to construct such trees is by connecting the heads in a lexicalized constituent parse. However, there are spanning trees that cannot be constructed in this way. Syntactic constituents are *contiguous spans*. In a spanning tree constructed from a lexicalized constituent parse, the head h of any constituent that spans the nodes from i to j must have a path to every node in this span. This property is known as **projectivity**, and projective dependency parses are a restricted class of spanning trees. Informally, projectivity means that “crossing edges” are prohibited. The formal definition follows:

	% non-projective edges	% non-projective sentences
Czech	1.86%	22.42%
English	0.39%	7.63%
German	2.33%	28.19%

Table 11.1: Frequency of non-projective dependencies in three languages (Kuhlmann and Nivre, 2010)

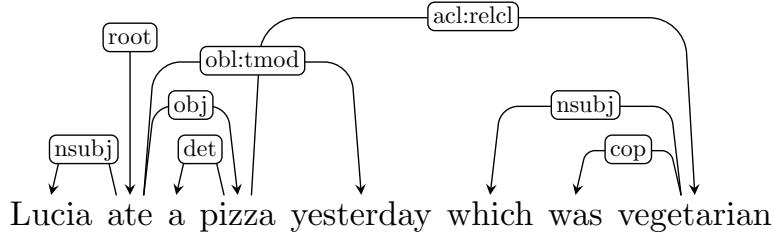


Figure 11.5: An example of a non-projective dependency parse. The “crossing edge” arises from the relative clause *which was vegetarian* and the oblique temporal modifier *yesterday*.

6084 **Definition 2** (Projectivity). *An edge from i to j is projective iff all k between i and j are descendants of i . A dependency parse is projective iff all its edges are projective.*

6086 Figure 11.5 gives an example of a non-projective dependency graph in English. This
6087 dependency graph does not correspond to any constituent parse. As shown in Table 11.1,
6088 non-projectivity is more common in languages such as Czech and German. Even though
6089 relatively few dependencies are non-projective in these languages, many sentences have
6090 at least one such dependency. As we will soon see, projectivity has important algorithmic
6091 consequences.

6092 11.2 Graph-based dependency parsing

6093 Let $\mathbf{y} = \{(i \xrightarrow{r} j)\}$ represent a dependency graph, in which each edge is a relation r from
6094 head word $i \in \{1, 2, \dots, M, \text{ROOT}\}$ to modifier $j \in \{1, 2, \dots, M\}$. The special node ROOT
6095 indicates the root of the graph, and M is the length of the input $|\mathbf{w}|$. Given a scoring
6096 function $\Psi(\mathbf{y}, \mathbf{w}; \theta)$, the optimal parse is,

$$\hat{\mathbf{y}} = \underset{\mathbf{y} \in \mathcal{Y}(\mathbf{w})}{\operatorname{argmax}} \Psi(\mathbf{y}, \mathbf{w}; \theta), \quad [11.1]$$

6097 where $\mathcal{Y}(\mathbf{w})$ is the set of valid dependency parses on the input \mathbf{w} . As usual, the number
6098 of possible labels $|\mathcal{Y}(\mathbf{w})|$ is exponential in the length of the input (Wu and Chao, 2004).

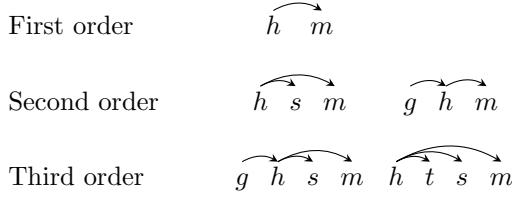


Figure 11.6: Feature templates for higher-order dependency parsing

6099 Algorithms that search over this space of possible graphs are known as **graph-based de-**
 6100 **pendency parsers.**

In sequence labeling and constituent parsing, it was possible to search efficiently over an exponential space by choosing a feature function that decomposes into a sum of local feature vectors. A similar approach is possible for dependency parsing, by requiring the scoring function to decompose across dependency arcs:

$$\Psi(\mathbf{y}, \mathbf{w}; \boldsymbol{\theta}) = \sum_{i \xrightarrow{r} j \in \mathbf{y}} \psi(i \xrightarrow{r} j, \mathbf{w}; \boldsymbol{\theta}). \quad [11.2]$$

6101 Dependency parsers that operate under this assumption are known as **arc-factored**, since
 6102 the score of a graph is the product of the scores of all arcs.

Higher-order dependency parsing The arc-factored decomposition can be relaxed to allow higher-order dependencies. In **second-order dependency parsing**, the scoring function may include grandparents and siblings, as shown by the templates in Figure 11.6. The scoring function is,

$$\begin{aligned} \Psi(\mathbf{y}, \mathbf{w}; \boldsymbol{\theta}) = & \sum_{i \xrightarrow{r} j \in \mathbf{y}} \psi_{\text{parent}}(i \xrightarrow{r} j, \mathbf{w}; \boldsymbol{\theta}) \\ & + \sum_{k \xrightarrow{r'} i \in \mathbf{y}} \psi_{\text{grandparent}}(i \xrightarrow{r} j, k, r', \mathbf{w}; \boldsymbol{\theta}) \\ & + \sum_{\substack{i \xrightarrow{r'} s \in \mathbf{y} \\ s \neq j}} \psi_{\text{ sibling}}(i \xrightarrow{r} j, s, r', \mathbf{w}; \boldsymbol{\theta}). \end{aligned} \quad [11.3]$$

6103 The top line scores computes a scoring function that includes the grandparent k ; the
 6104 bottom line computes a scoring function for each sibling s . For projective dependency
 6105 graphs, there are efficient algorithms for second-order and third-order dependency pars-
 6106 ing (Eisner, 1996; McDonald and Pereira, 2006; Koo and Collins, 2010); for non-projective
 6107 dependency graphs, second-order dependency parsing is NP-hard (McDonald and Pereira,
 6108 2006). The specific algorithms are discussed in the next section.

6109 **11.2.1 Graph-based parsing algorithms**

6110 The distinction between projective and non-projective dependency trees (§ 11.1.3) plays
 6111 a key role in the choice of algorithms. Because projective dependency trees are closely
 6112 related to (and can be derived from) lexicalized constituent trees, lexicalized parsing al-
 6113 gorithms can be applied directly. For the more general problem of parsing to arbitrary
 6114 spanning trees, a different class of algorithms is required. In both cases, arc-factored de-
 6115 pendency parsing relies on precomputing the scores $\psi(i \xrightarrow{r} j, w; \theta)$ for each potential
 6116 edge. There are $\mathcal{O}(M^2 R)$ such scores, where M is the length of the input and R is the
 6117 number of dependency relation types, and this is a lower bound on the time and space
 6118 complexity of any exact algorithm for arc-factored dependency parsing.

6119 **Projective dependency parsing**

6120 Any lexicalized constituency tree can be converted into a projective dependency tree by
 6121 creating arcs between the heads of constituents and their parents, so any algorithm for
 6122 lexicalized constituent parsing can be converted into an algorithm for projective depen-
 6123 dency parsing, by converting arc scores into scores for lexicalized productions. As noted
 6124 in § 10.5.2, there are cubic time algorithms for lexicalized constituent parsing, which are
 6125 extensions of the CKY algorithm. Therefore, arc-factored projective dependency parsing
 6126 can be performed in cubic time in the length of the input.

6127 Second-order projective dependency parsing can also be performed in cubic time, with
 6128 minimal modifications to the lexicalized parsing algorithm (Eisner, 1996). It is possible to
 6129 go even further, to **third-order dependency parsing**, in which the scoring function may
 6130 consider great-grandparents, grand-siblings, and “tri-siblings”, as shown in Figure 11.6.
 6131 Third-order dependency parsing can be performed in $\mathcal{O}(M^4)$ time, which can be made
 6132 practical through the use of pruning to eliminate unlikely edges (Koo and Collins, 2010).

6133 **Non-projective dependency parsing**

6134 In non-projective dependency parsing, the goal is to identify the highest-scoring span-
 6135 ning tree over the words in the sentence. The arc-factored assumption ensures that the
 6136 score for each spanning tree will be computed as a sum over scores for the edges, which
 6137 are precomputed. Based on these scores, we build a weighted connected graph. Arc-
 6138 factored non-projective dependency parsing is then equivalent to finding the spanning
 6139 tree that achieves the maximum total score, $\Psi(y, w) = \sum_{i \xrightarrow{r} j \in y} \psi(i \xrightarrow{r} j, w)$. The **Chu-**
 6140 **Liu-Edmonds algorithm** (Chu and Liu, 1965; Edmonds, 1967) computes this **maximum**
 6141 **directed spanning tree** efficiently. It does this by first identifying the best incoming edge
 6142 $i \xrightarrow{r} j$ for each vertex j . If the resulting graph does not contain cycles, it is the maxi-
 6143 mum spanning tree. If there is a cycle, it is collapsed into a super-vertex, whose incoming
 6144 and outgoing edges are based on the edges to the vertices in the cycle. The algorithm is

6145 then applied recursively to the resulting graph, and process repeats until a graph without
 6146 cycles is obtained.

6147 The time complexity of identifying the best incoming edge for each vertex is $\mathcal{O}(M^2R)$,
 6148 where M is the length of the input and R is the number of relations; in the worst case, the
 6149 number of cycles is $\mathcal{O}(M)$. Therefore, the complexity of the Chu-Liu-Edmonds algorithm
 6150 is $\mathcal{O}(M^3R)$. This complexity can be reduced to $\mathcal{O}(M^2N)$ by storing the edge scores in a
 6151 Fibonacci heap (Gabow et al., 1986). For more detail on graph-based parsing algorithms,
 6152 see Eisner (1997) and Kübler et al. (2009).

6153 **Higher-order non-projective dependency parsing** Given the tractability of higher-order
 6154 projective dependency parsing, you may be surprised to learn that non-projective second-
 6155 order dependency parsing is NP-Hard. This can be proved by reduction from the vertex
 6156 cover problem (Neuhaus and Bröker, 1997). A heuristic solution is to do projective pars-
 6157 ing first, and then post-process the projective dependency parse to add non-projective
 6158 edges (Nivre and Nilsson, 2005). More recent work has applied techniques for approxi-
 6159 mate inference in graphical models, including belief propagation (Smith and Eisner, 2008),
 6160 integer linear programming (Martins et al., 2009), variational inference (Martins et al.,
 6161 2010), and Markov Chain Monte Carlo (Zhang et al., 2014).

6162 11.2.2 Computing scores for dependency arcs

The arc-factored scoring function $\psi(i \xrightarrow{r} j, \mathbf{w}; \boldsymbol{\theta})$ can be defined in several ways:

$$\text{Linear} \quad \psi(i \xrightarrow{r} j, \mathbf{w}; \boldsymbol{\theta}) = \boldsymbol{\theta} \cdot \mathbf{f}(i \xrightarrow{r} j, \mathbf{w}) \quad [11.4]$$

$$\text{Neural} \quad \psi(i \xrightarrow{r} j, \mathbf{w}; \boldsymbol{\theta}) = \text{Feedforward}([\mathbf{u}_{w_i}; \mathbf{u}_{w_j}]; \boldsymbol{\theta}) \quad [11.5]$$

$$\text{Generative} \quad \psi(i \xrightarrow{r} j, \mathbf{w}; \boldsymbol{\theta}) = \log p(w_j, r | w_i). \quad [11.6]$$

6163 Linear feature-based arc scores

6164 Linear models for dependency parsing incorporate many of the same features used in
 6165 sequence labeling and discriminative constituent parsing. These include:

- 6166 • the length and direction of the arc;
- 6167 • the words w_i and w_j linked by the dependency relation;
- 6168 • the prefixes, suffixes, and parts-of-speech of these words;
- 6169 • the neighbors of the dependency arc, $w_{i-1}, w_{i+1}, w_{j-1}, w_{j+1}$;
- 6170 • the prefixes, suffixes, and part-of-speech of these neighbor words.

6171 Each of these features can be conjoined with the dependency edge label r . Note that
 6172 features in an arc-factored parser can refer to words other than w_i and w_j . The restriction
 6173 is that the features consider only a single arc.

Bilexical features (e.g., *sushi* → *chopsticks*) are powerful but rare, so it is useful to augment them with coarse-grained alternatives, by “backing off” to the part-of-speech or affix. For example, the following features are created by backing off to part-of-speech tags in an unlabeled dependency parser:

$$\begin{aligned} f(3 \rightarrow 5, \text{we eat sushi with chopsticks}) = & \langle \text{sushi} \rightarrow \text{chopsticks}, \\ & \text{sushi} \rightarrow \text{NNS}, \\ & \text{NN} \rightarrow \text{chopsticks}, \\ & \text{NNS} \rightarrow \text{NN} \rangle. \end{aligned}$$

6174 Regularized discriminative learning algorithms can then trade off between features at
 6175 varying levels of detail. McDonald et al. (2005) take this approach as far as *tetralexical*
 6176 features (e.g., $(w_i, w_{i+1}, w_{j-1}, w_j)$). Such features help to avoid choosing arcs that are un-
 6177 likely due to the intervening words: for example, there is unlikely to be an edge between
 6178 two nouns if the intervening span contains a verb. A large list of first and second-order
 6179 features is provided by Bohnet (2010), who uses a hashing function to store these features
 6180 efficiently.

6181 Neural arc scores

Given vector representations \mathbf{x}_i for each word w_i in the input, a set of arc scores can be computed from a feedforward neural network:

$$\psi(i \xrightarrow{r} j, \mathbf{w}; \boldsymbol{\theta}) = \text{FeedForward}([\mathbf{x}_i; \mathbf{x}_j]; \boldsymbol{\theta}_r), \quad [11.7]$$

where unique weights $\boldsymbol{\theta}_r$ are available for each arc type (Pei et al., 2015; Kiperwasser and Goldberg, 2016). Kiperwasser and Goldberg (2016) use a feedforward network with a single hidden layer,

$$\mathbf{z} = g(\boldsymbol{\Theta}_r[\mathbf{x}_i; \mathbf{x}_j] + \mathbf{b}_r^{(z)}) \quad [11.8]$$

$$\psi(i \xrightarrow{r} j) = \boldsymbol{\beta}_r \mathbf{z} + \mathbf{b}_r^{(y)}, \quad [11.9]$$

6182 where $\boldsymbol{\Theta}_r$ is a matrix, $\boldsymbol{\beta}_r$ is a vector, each b_r is a scalar, and the function g is an elementwise
 6183 tanh activation function.

6184 The vector \mathbf{x}_i can be set equal to the word embedding, which may be pre-trained or
 6185 learned by backpropagation (Pei et al., 2015). Alternatively, contextual information can
 6186 be incorporated by applying a bidirectional recurrent neural network across the input, as

described in § 7.6. The RNN hidden states at each word can be used as inputs to the arc scoring function (Kiperwasser and Goldberg, 2016).

Feature-based arc scores are computationally expensive, due to the costs of storing and searching a huge table of weights. Neural arc scores can be viewed as a compact solution to this problem. Rather than working in the space of tuples of lexical features, the hidden layers of a feedforward network can be viewed as implicitly computing feature combinations, with each layer of the network evaluating progressively more words. An early paper on neural dependency parsing showed substantial speed improvements at test time, while also providing higher accuracy than feature-based models (Chen and Manning, 2014).

Probabilistic arc scores

If each arc score is equal to the log probability $\log p(w_j, r \mid w_i)$, then the sum of scores gives the log probability of the sentence and arc labels, by the chain rule. For example, consider the unlabeled parse of *we eat sushi with rice*,

$$\mathbf{y} = \{(ROOT, 2), (2, 1), (2, 3), (3, 5), (5, 4)\} \quad [11.10]$$

$$\log p(\mathbf{w} \mid \mathbf{y}) = \sum_{(i \rightarrow j) \in \mathbf{y}} \log p(w_j \mid w_i) \quad [11.11]$$

$$\begin{aligned} &= \log p(eat \mid ROOT) + \log p(we \mid eat) + \log p(sushi \mid eat) \\ &\quad + \log p(rice \mid sushi) + \log p(with \mid rice). \end{aligned} \quad [11.12]$$

Probabilistic generative models are used in combination with expectation-maximization (chapter 5) for unsupervised dependency parsing (Klein and Manning, 2004).

11.2.3 Learning

Having formulated graph-based dependency parsing as a structure prediction problem, we can apply similar learning algorithms to those used in sequence labeling. Given a loss function $\ell(\boldsymbol{\theta}; \mathbf{w}^{(i)}, \mathbf{y}^{(i)})$, we can compute gradient-based updates to the parameters. For a model with feature-based arc scores and a perceptron loss, we obtain the usual structured perceptron update,

$$\hat{\mathbf{y}} = \operatorname{argmax}_{\mathbf{y}' \in \mathcal{Y}(\mathbf{w})} \boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{w}, \mathbf{y}') \quad [11.13]$$

$$\boldsymbol{\theta} = \boldsymbol{\theta} + \mathbf{f}(\mathbf{w}, \mathbf{y}) - \mathbf{f}(\mathbf{w}, \hat{\mathbf{y}}) \quad [11.14]$$

In this case, the argmax requires a maximization over all dependency trees for the sentence, which can be computed using the algorithms described in § 11.2.1. We can apply all the usual tricks from § 2.3: weight averaging, a large margin objective, and regularization. McDonald et al. (2005) were the first to treat dependency parsing as a structure

6205 prediction problem, using MIRA, an online margin-based learning algorithm. Neural arc
 6206 scores can be learned in the same way, backpropagating from a margin loss to updates on
 6207 the feedforward network that computes the score for each edge.

A conditional random field for arc-factored dependency parsing is built on the probability model,

$$p(\mathbf{y} \mid \mathbf{w}) = \frac{\exp \sum_{i \xrightarrow{r} j \in \mathbf{y}} \psi(i \xrightarrow{r} j, \mathbf{w}; \boldsymbol{\theta})}{\sum_{\mathbf{y}' \in \mathcal{Y}(\mathbf{w})} \exp \sum_{i \xrightarrow{r} j \in \mathbf{y}'} \psi(i \xrightarrow{r} j, \mathbf{w}; \boldsymbol{\theta})} \quad [11.15]$$

6208 Such a model is trained to minimize the negative log conditional-likelihood. Just as in
 6209 CRF sequence models (§ 7.5.3) and the logistic regression classifier (§ 2.5), the gradients
 6210 involve marginal probabilities $p(i \xrightarrow{r} j \mid \mathbf{w}; \boldsymbol{\theta})$, which in this case are probabilities over
 6211 individual dependencies. In arc-factored models, these probabilities can be computed
 6212 in polynomial time. For projective dependency trees, the marginal probabilities can be
 6213 computed in cubic time, using a variant of the inside-outside algorithm (Lari and Young,
 6214 1990). For non-projective dependency parsing, marginals can also be computed in cubic
 6215 time, using the **matrix-tree theorem** (Koo et al., 2007; McDonald et al., 2007; Smith and
 6216 Smith, 2007). Details of these methods are described by Kübler et al. (2009).

6217 11.3 Transition-based dependency parsing

6218 Graph-based dependency parsing offers exact inference, meaning that it is possible to re-
 6219 cover the best-scoring parse for any given model. But this comes at a price: the scoring
 6220 function is required to decompose into local parts — in the case of non-projective parsing,
 6221 these parts are restricted to individual arcs. These limitations are felt more keenly in de-
 6222 pendency parsing than in sequence labeling, because second-order dependency features
 6223 are critical to correctly identify some types of attachments. For example, prepositional
 6224 phrase attachment depends on the attachment point, the object of the preposition, and
 6225 the preposition itself; arc-factored scores cannot account for all three of these features si-
 6226 multaneously. Graph-based dependency parsing may also be criticized on the basis of
 6227 intuitions about human language processing: people read and listen to sentences *sequen-*
 6228 *tially*, incrementally building mental models of the sentence structure and meaning before
 6229 getting to the end (Jurafsky, 1996). This seems hard to reconcile with graph-based algo-
 6230 rithms, which perform bottom-up operations on the entire sentence, requiring the parser
 6231 to keep every word in memory. Finally, from a practical perspective, graph-based depen-
 6232 dency parsing is relatively slow, running in cubic time in the length of the input.

6233 Transition-based algorithms address all three of these objections. They work by mov-
 6234 ing through the sentence sequentially, while performing actions that incrementally up-
 6235 date a stored representation of what has been read thus far. As with the shift-reduce

parser from § 10.6.2, this representation consists of a stack, onto which parsing substructures can be pushed and popped. In shift-reduce, these substructures were constituents; in the transition systems that follow, they will be projective dependency trees over partial spans of the input.⁴ Parsing is complete when the input is consumed and there is only a single structure on the stack. The sequence of actions that led to the parse is known as the **derivation**. One problem with transition-based systems is that there may be multiple derivations for a single parse structure — a phenomenon known as **spurious ambiguity**.

11.3.1 Transition systems for dependency parsing

A **transition system** consists of a representation for describing configurations of the parser, and a set of transition actions, which manipulate the configuration. There are two main transition systems for dependency parsing: **arc-standard**, which is closely related to shift-reduce, and **arc-eager**, which adds an additional action that can simplify derivations (Abney and Johnson, 1991). In both cases, transitions are between **configurations** that are represented as triples, $C = (\sigma, \beta, A)$, where σ is the stack, β is the input buffer, and A is the list of arcs that have been created (Nivre, 2008). In the initial configuration,

$$C_{\text{initial}} = ([\text{ROOT}], \mathbf{w}, \emptyset), \quad [11.16]$$

indicating that the stack contains only the special node ROOT, the entire input is on the buffer, and the set of arcs is empty. An accepting configuration is,

$$C_{\text{accept}} = ([\text{ROOT}], \emptyset, A), \quad [11.17]$$

where the stack contains only ROOT, the buffer is empty, and the arcs A define a spanning tree over the input. The arc-standard and arc-eager systems define a set of transitions between configurations, which are capable of transforming an initial configuration into an accepting configuration. In both of these systems, the number of actions required to parse an input grows linearly in the length of the input, making transition-based parsing considerably more efficient than graph-based methods.

Arc-standard

The **arc-standard** transition system is closely related to shift-reduce, and to the LR algorithm that is used to parse programming languages (Aho et al., 2006). It includes the following classes of actions:

- SHIFT: move the first item from the input buffer on to the top of the stack,

$$(\sigma, i|\beta, A) \Rightarrow (\sigma|i, \beta, A), \quad [11.18]$$

⁴Transition systems also exist for non-projective dependency parsing (e.g., Nivre, 2008).

6264 where we write $i|\beta$ to indicate that i is the leftmost item in the input buffer, and $\sigma|i$
 6265 to indicate the result of pushing i on to stack σ .

- 6266 • ARC-LEFT: create a new left-facing arc of type r between the item on the top of the
 6267 stack and the first item in the input buffer. The head of this arc is j , which remains
 6268 at the front of the input buffer. The arc $j \xrightarrow{r} i$ is added to A . Formally,

$$(\sigma|i, j|\beta, A) \Rightarrow (\sigma, j|\beta, A \oplus j \xrightarrow{r} i), \quad [11.19]$$

6269 where r is the label of the dependency arc, and \oplus concatenates the new arc $j \xrightarrow{r} i$ to
 6270 the list A .

- 6271 • ARC-RIGHT: creates a new right-facing arc of type r between the item on the top of the
 6272 stack and the first item in the input buffer. The head of this arc is i , which is
 6273 “popped” from the stack and pushed to the front of the input buffer. The arc $i \xrightarrow{r} j$
 6274 is added to A . Formally,

$$(\sigma|i, j|\beta, A) \Rightarrow (\sigma, i|\beta, A \oplus i \xrightarrow{r} j), \quad [11.20]$$

6275 where again r is the label of the dependency arc.

6276 Each action has preconditions. The SHIFT action can be performed only when the buffer
 6277 has at least one element. The ARC-LEFT action cannot be performed when the root node
 6278 ROOT is on top of the stack, since this node must be the root of the entire tree. The ARC-
 6279 LEFT and ARC-RIGHT remove the modifier words from the stack (in the case of ARC-LEFT)
 6280 and from the buffer (in the case of ARC-RIGHT), so it is impossible for any word to have
 6281 more than one parent. Furthermore, the end state can only be reached when every word is
 6282 removed from the buffer and stack, so the set of arcs is guaranteed to constitute a spanning
 6283 tree. An example arc-standard derivation is shown in Table 11.2.

6284 Arc-eager dependency parsing

6285 In the arc-standard transition system, a word is completely removed from the parse once
 6286 it has been made the modifier in a dependency arc. At this time, any dependents of
 6287 this word must have already been identified. Right-branching structures are common in
 6288 English (and many other languages), with words often modified by units such as prepo-
 6289 sitional phrases to their right. In the arc-standard system, this means that we must first
 6290 shift all the units of the input onto the stack, and then work backwards, creating a series of
 6291 arcs, as occurs in Table 11.2. Note that the decision to shift *bagels* onto the stack guarantees
 6292 that the prepositional phrase *with lox* will attach to the noun phrase, and that this decision
 6293 must be made before the prepositional phrase is itself parsed. This has been argued to be
 6294 cognitively implausible (Abney and Johnson, 1991); from a computational perspective, it
 6295 means that a parser may need to look several steps ahead to make the correct decision.

σ	β	action	arc added to \mathcal{A}
1. [ROOT]	<i>they like bagels with lox</i>	SHIFT	
2. [ROOT, <i>they</i>]	<i>like bagels with lox</i>	ARC-LEFT	(<i>they</i> \leftarrow <i>like</i>)
3. [ROOT]	<i>like bagels with lox</i>	SHIFT	
4. [ROOT, <i>like</i>]	<i>bagels with lox</i>	SHIFT	
5. [ROOT, <i>like</i> , <i>bagels</i>]	<i>with lox</i>	SHIFT	
6. [ROOT, <i>like</i> , <i>bagels</i> , <i>with</i>]	<i>lox</i>	ARC-LEFT	(<i>with</i> \leftarrow <i>lox</i>)
7. [ROOT, <i>like</i> , <i>bagels</i>]	<i>lox</i>	ARC-RIGHT	(<i>bagels</i> \rightarrow <i>lox</i>)
8. [ROOT, <i>like</i>]	<i>bagels</i>	ARC-RIGHT	(<i>like</i> \rightarrow <i>bagels</i>)
9. [ROOT]	<i>like</i>	ARC-RIGHT	(ROOT \rightarrow <i>like</i>)
10. [ROOT]	\emptyset	DONE	

Table 11.2: Arc-standard derivation of the unlabeled dependency parse for the input *they like bagels with lox*.

6296 **Arc-eager dependency parsing** changes the ARC-RIGHT action so that right depen-
 6297 dents can be attached before all of their dependents have been found. Rather than re-
 6298 moving the modifier from both the buffer and stack, the ARC-RIGHT action pushes the
 6299 modifier on to the stack, on top of the head. Because the stack can now contain elements
 6300 that already have parents in the partial dependency graph, two additional changes are
 6301 necessary:

- 6302 • A precondition is required to ensure that the ARC-LEFT action cannot be applied
 6303 when the top element on the stack already has a parent in A .
 6304 • A new REDUCE action is introduced, which can remove elements from the stack if
 6305 they already have a parent in A :

$$(\sigma|i, \beta, A) \Rightarrow (\sigma, \beta, A). \quad [11.21]$$

6306 As a result of these changes, it is now possible to create the arc *like* \rightarrow *bagels* before parsing
 6307 the prepositional phrase *with lox*. Furthermore, this action does not imply a decision about
 6308 whether the prepositional phrase will attach to the noun or verb. Noun attachment is
 6309 chosen in the parse in Table 11.3, but verb attachment could be achieved by applying the
 6310 REDUCE action at step 5 or 7.

6311 Projectivity

6312 The arc-standard and arc-eager transition systems are guaranteed to produce projective
 6313 dependency trees, because all arcs are between the word at the top of the stack and the

σ	β	action	arc added to \mathcal{A}
1. [ROOT]	<i>they like bagels with lox</i>	SHIFT	
2. [ROOT, <i>they</i>]	<i>like bagels with lox</i>	ARC-LEFT	(<i>they</i> \leftarrow <i>like</i>)
3. [ROOT]	<i>like bagels with lox</i>	ARC-RIGHT	(ROOT \rightarrow <i>like</i>)
4. [ROOT, <i>like</i>]	<i>bagels with lox</i>	ARC-RIGHT	(<i>like</i> \rightarrow <i>bagels</i>)
5. [ROOT, <i>like</i> , <i>bagels</i>]	<i>with lox</i>	SHIFT	
6. [ROOT, <i>like</i> , <i>bagels</i> , <i>with</i>]	<i>lox</i>	ARC-LEFT	(<i>with</i> \leftarrow <i>lox</i>)
7. [ROOT, <i>like</i> , <i>bagels</i>]	<i>lox</i>	ARC-RIGHT	(<i>bagels</i> \rightarrow <i>lox</i>)
8. [ROOT, <i>like</i> , <i>bagels</i> , <i>lox</i>]	\emptyset	REDUCE	
9. [ROOT, <i>like</i> , <i>bagels</i>]	\emptyset	REDUCE	
10. [ROOT, <i>like</i>]	\emptyset	REDUCE	
11. [ROOT]	\emptyset	DONE	

Table 11.3: Arc-eager derivation of the unlabeled dependency parse for the input *they like bagels with lox*.

6314 left-most edge of the buffer (Nivre, 2008). Non-projective transition systems can be con-
 6315 structed by adding actions that create arcs with words that are second or third in the
 6316 stack (Attardi, 2006), or by adopting an alternative configuration structure, which main-
 6317 tains a list of all words that do not yet have heads (Covington, 2001). In **pseudo-projective**
 6318 **dependency parsing**, a projective dependency parse is generated first, and then a set of
 6319 graph transformation techniques are applied, producing non-projective edges (Nivre and
 6320 Nilsson, 2005).

6321 Beam search

6322 In “greedy” transition-based parsing, the parser tries to make the best decision at each
 6323 configuration. This can lead to search errors, when an early decision locks the parser into
 6324 a poor derivation. For example, in Table 11.2, if ARC-RIGHT were chosen at step 4, then
 6325 the parser would later be forced to attach the prepositional phrase *with lox* to the verb
 6326 *likes*. Note that the *likes* \rightarrow *bagels* arc is indeed part of the correct dependency parse, but
 6327 the arc-standard transition system requires it to be created later in the derivation.

Beam search is a general technique for ameliorating search errors in incremental de-
 coding.⁵ While searching, the algorithm maintains a set of partially-complete hypotheses,
 called a beam. At step t of the derivation, there is a set of k hypotheses, each of which

⁵Beam search is used throughout natural language processing, and beyond. In this text, it appears again in coreference resolution (§ 15.2.4) and machine translation (§ 18.4).

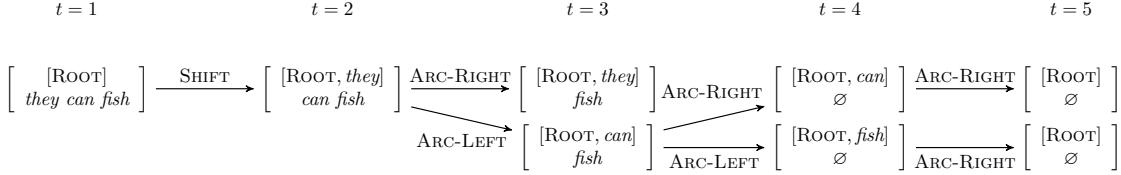


Figure 11.7: Beam search for unlabeled dependency parsing, with beam size $K = 2$. The arc lists for each configuration are not shown, but can be computed from the transitions.

includes a score $s_t^{(k)}$ and a set of dependency arcs $A_t^{(k)}$:

$$h_t^{(k)} = (s_t^{(k)}, A_t^{(k)}) \quad [11.22]$$

6328 Each hypothesis is then “expanded” by considering the set of all valid actions from the
 6329 current configuration $c_t^{(k)}$, written $\mathcal{A}(c_t^{(k)})$. This yields a large set of new hypotheses. For
 6330 each action $a \in \mathcal{A}(c_t^{(k)})$, we score the new hypothesis $A_t^{(k)} \oplus a$. The top k hypotheses
 6331 by this scoring metric are kept, and parsing proceeds to the next step (Zhang and Clark,
 6332 2008). Note that beam search requires a scoring function for action *sequences*, rather than
 6333 individual actions. This issue will be revisited in the next section.

6334 Figure 11.7 shows the application of beam search to dependency parsing, with a beam
 6335 size of $K = 2$. For the first transition, the only valid action is SHIFT, so there is only
 6336 one possible configuration at $t = 2$. From this configuration, there are three possible
 6337 actions. The two best scoring actions are ARC-RIGHT and ARC-LEFT, and so the resulting
 6338 hypotheses from these actions are on the beam at $t = 3$. From these configurations, there
 6339 are three possible actions each, but the best two are expansions of the bottom hypothesis
 6340 at $t = 3$. Parsing continues until $t = 5$, at which point both hypotheses reach an accepting
 6341 state. The best-scoring hypothesis is then selected as the parse.

6342 11.3.2 Scoring functions for transition-based parsers

Transition-based parsing requires selecting a series of actions. In greedy transition-based
 parsing, this can be done by training a classifier,

$$\hat{a} = \underset{a \in \mathcal{A}(c)}{\operatorname{argmax}} \Psi(a, c, \mathbf{w}; \boldsymbol{\theta}), \quad [11.23]$$

6343 where $\mathcal{A}(c)$ is the set of admissible actions in the current configuration c , \mathbf{w} is the input,
 6344 and Ψ is a scoring function with parameters $\boldsymbol{\theta}$ (Yamada and Matsumoto, 2003).

6345 A feature-based score can be computed, $\Psi(a, c, \mathbf{w}) = \boldsymbol{\theta} \cdot \mathbf{f}(a, c, \mathbf{w})$, using features that
 6346 may consider any aspect of the current configuration and input sequence. Typical features
 6347 for transition-based dependency parsing include: the word and part-of-speech of the top

6348 element on the stack; the word and part-of-speech of the first, second, and third elements
 6349 on the input buffer; pairs and triples of words and parts-of-speech from the top of the
 6350 stack and the front of the buffer; the distance (in tokens) between the element on the top
 6351 of the stack and the element in the front of the input buffer; the number of modifiers of
 6352 each of these elements; and higher-order dependency features as described above in the
 6353 section on graph-based dependency parsing (see, e.g., Zhang and Nivre, 2011).

6354 Parse actions can also be scored by neural networks. For example, Chen and Manning
 6355 (2014) build a feedforward network in which the input layer consists of the concatenation
 6356 of embeddings of several words and tags:

- 6357 • the top three words on the stack, and the first three words on the buffer;
- 6358 • the first and second leftmost and rightmost children (dependents) of the top two
 6359 words on the stack;
- 6360 • the leftmost and right most grandchildren of the top two words on the stack;
- 6361 • embeddings of the part-of-speech tags of these words.

Let us call this base layer $\mathbf{x}(c, \mathbf{w})$, defined as,

$$c = (\sigma, \beta, A)$$

$$\mathbf{x}(c, \mathbf{w}) = [\mathbf{v}_{w_{\sigma_1}}, \mathbf{v}_{t_{\sigma_1}} \mathbf{v}_{w_{\sigma_2}}, \mathbf{v}_{t_{\sigma_2}}, \mathbf{v}_{w_{\sigma_3}}, \mathbf{v}_{t_{\sigma_3}}, \mathbf{v}_{w_{\beta_1}}, \mathbf{v}_{t_{\beta_1}}, \mathbf{v}_{w_{\beta_2}}, \mathbf{v}_{t_{\beta_2}}, \dots],$$

where $\mathbf{v}_{w_{\sigma_1}}$ is the embedding of the first word on the stack, $\mathbf{v}_{t_{\beta_2}}$ is the embedding of the part-of-speech tag of the second word on the buffer, and so on. Given this base encoding of the parser state, the score for the set of possible actions is computed through a feedforward network,

$$z = g(\Theta^{(x \rightarrow z)} \mathbf{x}(c, \mathbf{w})) \quad [11.24]$$

$$\psi(a, c, \mathbf{w}; \theta) = \Theta_a^{(z \rightarrow y)} z, \quad [11.25]$$

6362 where the vector z plays the same role as the features $f(a, c, \mathbf{w})$, but is a learned represen-
 6363 tation. Chen and Manning (2014) use a cubic elementwise activation function, $g(x) = x^3$,
 6364 so that the hidden layer models products across all triples of input features. The learning
 6365 algorithm updates the embeddings as well as the parameters of the feedforward network.

6366 11.3.3 Learning to parse

6367 Transition-based dependency parsing suffers from a mismatch between the supervision,
 6368 which comes in the form of dependency trees, and the classifier's prediction space, which
 6369 is a set of parsing actions. One solution is to create new training data by converting parse
 6370 trees into action sequences; another is to derive supervision directly from the parser's
 6371 performance.

6372 **Oracle-based training**

6373 A transition system can be viewed as a function from action sequences (derivations) to
 6374 parse trees. The inverse of this function is a mapping from parse trees to derivations,
 6375 which is called an **oracle**. For the arc-standard and arc-eager parsing system, an oracle can
 6376 be computed in linear time in the length of the derivation (Kübler et al., 2009, page 32).
 6377 Both the arc-standard and arc-eager transition systems suffer from spurious ambiguity:
 6378 there exist dependency parses for which multiple derivations are possible, such as $1 \leftarrow 2 \rightarrow 3$. The oracle must choose between these different derivations. For example, the
 6380 algorithm described by Kübler et al. (2009) would first create the left arc ($1 \leftarrow 2$), and then
 6381 create the right arc, $(1 \leftarrow 2) \rightarrow 3$; another oracle might begin by shifting twice, resulting
 6382 in the derivation $1 \leftarrow (2 \rightarrow 3)$.

Given such an oracle, a dependency treebank can be converted into a set of oracle action sequences $\{A^{(i)}\}_{i=1}^N$. The parser can be trained by stepping through the oracle action sequences, and optimizing on an classification-based objective that rewards selecting the oracle action. For transition-based dependency parsing, maximum conditional likelihood is a typical choice (Chen and Manning, 2014; Dyer et al., 2015):

$$p(a | c, \mathbf{w}) = \frac{\exp \Psi(a, c, \mathbf{w}; \theta)}{\sum_{a' \in \mathcal{A}(c)} \exp \Psi(a', c, \mathbf{w}; \theta)} \quad [11.26]$$

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^N \sum_{t=1}^{|A^{(i)}|} \log p(a_t^{(i)} | c_t^{(i)}, \mathbf{w}), \quad [11.27]$$

6383 where $|A^{(i)}|$ is the length of the action sequence $A^{(i)}$.

6384 Recall that beam search requires a scoring function for action sequences. Such a score
 6385 can be obtained by adding the log-likelihoods (or hinge losses) across all actions in the
 6386 sequence (Chen and Manning, 2014).

6387 **Global objectives**

6388 The objective in Equation 11.27 is **locally-normalized**: it is the product of normalized
 6389 probabilities over individual actions. A similar characterization could be made of non-
 6390 probabilistic algorithms in which hinge-loss objectives are summed over individual ac-
 6391 tions. In either case, training on individual actions can be sub-optimal with respect to
 6392 global performance, due to the **label bias problem** (Lafferty et al., 2001; Andor et al.,
 6393 2016).

6394 As a stylized example, suppose that a given configuration appears 100 times in the
 6395 training data, with action a_1 as the oracle action in 51 cases, and a_2 as the oracle action in
 6396 the other 49 cases. However, in cases where a_2 is correct, choosing a_1 results in a cascade
 6397 of subsequent errors, while in cases where a_1 is correct, choosing a_2 results in only a single

6398 error. A classifier that is trained on a local objective function will learn to always choose
 6399 a_1 , but choosing a_2 would minimize the overall number of errors.

6400 This observation motivates a global objective, such as the globally-normalized condi-
 6401 tional likelihood,

$$p(A^{(i)} | \mathbf{w}; \boldsymbol{\theta}) = \frac{\exp \sum_{t=1}^{|A^{(i)}|} \Psi(a_t^{(i)}, c_t^{(i)}, \mathbf{w})}{\sum_{A' \in \mathbb{A}(\mathbf{w})} \exp \sum_{t=1}^{|A'|} \Psi(a'_t, c'_t, \mathbf{w})}, \quad [11.28]$$

where the denominator sums over the set of all possible action sequences, $\mathbb{A}(\mathbf{w})$.⁶ In the conditional random field model for sequence labeling (§ 7.5.3), it was possible to compute this sum explicitly, using dynamic programming. In transition-based parsing, this is not possible. However, the sum can be approximated using beam search,

$$\sum_{A' \in \mathbb{A}(\mathbf{w})} \exp \sum_{t=1}^{|A'|} \Psi(a'_t, c'_t, \mathbf{w}) \approx \sum_{k=1}^K \exp \sum_{t=1}^{|A^{(k)}|} \Psi(a_t^{(k)}, c_t^{(k)}, \mathbf{w}), \quad [11.29]$$

where $A^{(k)}$ is an action sequence on a beam of size K . This gives rise to the following loss function,

$$L(\boldsymbol{\theta}) = - \sum_{t=1}^{|A^{(i)}|} \Psi(a_t^{(i)}, c_t^{(i)}, \mathbf{w}) + \log \sum_{k=1}^K \exp \sum_{t=1}^{|A^{(k)}|} \Psi(a_t^{(k)}, c_t^{(k)}, \mathbf{w}). \quad [11.30]$$

6402 The derivatives of this loss involve expectations with respect to a probability distribution
 6403 over action sequences on the beam.

6404 *Early update and the incremental perceptron

6405 When learning in the context of beam search, the goal is to learn a decision function so that
 6406 the gold dependency parse is always reachable from at least one of the partial derivations
 6407 on the beam. (The combination of a transition system (such as beam search) and a scoring
 6408 function for actions is known as a **policy**.) To achieve this, we can make an **early update**
 6409 as soon as the oracle action sequence “falls off” the beam, even before a complete analysis
 6410 is available (Collins and Roark, 2004; Daumé III and Marcu, 2005). The loss can be based
 6411 on the best-scoring hypothesis on the beam, or the sum of all hypotheses (Huang et al.,
 6412 2012).

6413 For example, consider the beam search in Figure 11.7. In the correct parse, *fish* is the
 6414 head of dependency arcs to both of the other two words. In the arc-standard system,

⁶Andor et al. (2016) prove that the set of globally-normalized conditional distributions is a strict superset of the set of locally-normalized conditional distributions, and that globally-normalized conditional models are therefore strictly more expressive.

6415 this can be achieved only by using SHIFT for the first two actions. At $t = 3$, the oracle
 6416 action sequence has fallen off the beam. The parser should therefore stop, and update the
 6417 parameters by the gradient $\frac{\partial}{\partial \theta} L(A_{1:3}^{(i)}, \{A_{1:3}^{(k)}\}; \theta)$, where $A_{1:3}^{(i)}$ is the first three actions of the
 6418 oracle sequence, and $\{A_{1:3}^{(k)}\}$ is the beam.

6419 This integration of incremental search and learning was first developed in the **incre-
 6420 mental perceptron** (Collins and Roark, 2004). This method updates the parameters with
 6421 respect to a hinge loss, which compares the top-scoring hypothesis and the gold action
 6422 sequence, up to the current point t . Several improvements to this basic protocol are pos-
 6423 sible:

- 6424 • As noted earlier, the gold dependency parse can be derived by multiple action se-
 6425 quences. Rather than checking for the presence of a single oracle action sequence on
 6426 the beam, we can check if the gold dependency parse is *reachable* from the current
 6427 beam, using a **dynamic oracle** (Goldberg and Nivre, 2012).
- 6428 • By maximizing the score of the gold action sequence, we are training a decision
 6429 function to find the correct action given the gold context. But in reality, the parser
 6430 will make errors, and the parser is not trained to find the best action given a context
 6431 that may not itself be optimal. This issue is addressed by various generalizations of
 6432 incremental perceptron, known as **learning to search** (Daumé III et al., 2009). Some
 6433 of these methods are discussed in chapter 15.

6434 11.4 Applications

6435 Dependency parsing is used in many real-world applications: any time you want to know
 6436 about pairs of words which might not be adjacent, you can use dependency arcs instead
 6437 of regular expression search patterns. For example, you may want to match strings like
 6438 *delicious pastries*, *delicious French pastries*, and *the pastries are delicious*.

6439 It is possible to search the Google n -grams corpus by dependency edges, finding the
 6440 trend in how often a dependency edge appears over time. For example, we might be inter-
 6441 ested in knowing when people started talking about *writing code*, but we also want *write*
 6442 *some code*, *write good code*, *write all the code*, etc. The result of a search on the dependency
 6443 edge *write* → *code* is shown in Figure 11.8. This capability has been applied to research
 6444 in digital humanities, such as the analysis of gender in Shakespeare Muralidharan and
 6445 Hearst (2013).

A classic application of dependency parsing is **relation extraction**, which is described

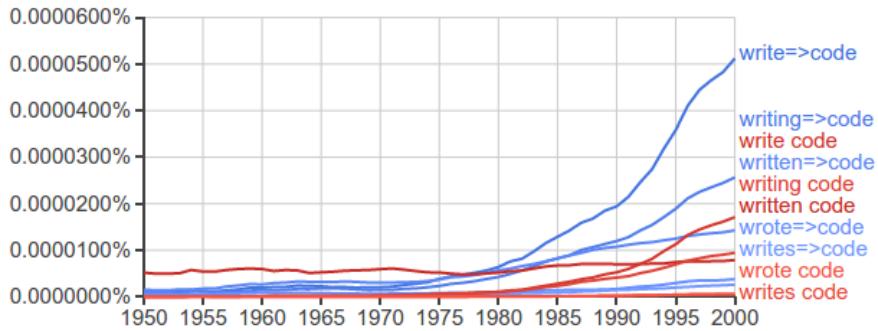


Figure 11.8: Google n-grams results for the bigram *write code* and the dependency arc *write => code* (and their morphological variants)

in chapter 17. The goal of relation extraction is to identify entity pairs, such as

(MELVILLE, MOBY-DICK)
 (TOLSTOY, WAR AND PEACE)
 (MARQUÉZ, 100 YEARS OF SOLITUDE)
 (SHAKESPEARE, A MIDSUMMER NIGHT'S DREAM),

6446 which stand in some relation to each other (in this case, the relation is authorship). Such
 6447 entity pairs are often referenced via consistent chains of dependency relations. Therefore,
 6448 dependency paths are often a useful feature in supervised systems which learn to detect
 6449 new instances of a relation, based on labeled examples of other instances of the same
 6450 relation type (Culotta and Sorensen, 2004; Fundel et al., 2007; Mintz et al., 2009).

6451 Cui et al. (2005) show how dependency parsing can improve automated question an-
 6452 swering. Suppose you receive the following query:

6453 (11.1) What percentage of the nation's cheese does Wisconsin produce?

6454 The corpus contains this sentence:

6455 (11.2) In Wisconsin, where farmers produce 28% of the nation's cheese, ...

6456 The location of *Wisconsin* in the surface form of this string makes it a poor match for the
 6457 query. However, in the dependency graph, there is an edge from *produce* to *Wisconsin* in
 6458 both the question and the potential answer, raising the likelihood that this span of text is
 6459 relevant to the question.

6460 A final example comes from sentiment analysis. As discussed in chapter 4, the polarity
 6461 of a sentence can be reversed by negation, e.g.

6462 (11.3) *There is no reason at all to believe the polluters will suddenly become reasonable.*

6463 By tracking the sentiment polarity through the dependency parse, we can better iden-
 6464 tify the overall polarity of the sentence, determining when key sentiment words are re-
 6465 versed (Wilson et al., 2005; Nakagawa et al., 2010).

6466 Additional resources

6467 More details on dependency grammar and parsing algorithms can be found in the manuscript
 6468 by Kübler et al. (2009). For a comprehensive but whimsical overview of graph-based de-
 6469 pendency parsing algorithms, see Eisner (1997). Jurafsky and Martin (2018) describe an
 6470 **agenda-based** version of beam search, in which the beam contains hypotheses of varying
 6471 lengths. New hypotheses are added to the beam only if their score is better than the worst
 6472 item currently on the beam. Another search algorithm for transition-based parsing is
 6473 **easy-first**, which abandons the left-to-right traversal order, and adds the highest-scoring
 6474 edges first, regardless of where they appear (Goldberg and Elhadad, 2010). Goldberg et al.
 6475 (2013) note that although transition-based methods can be implemented in linear time in
 6476 the length of the input, naïve implementations of beam search will require quadratic time,
 6477 due to the cost of copying each hypothesis when it is expanded on the beam. This issue
 6478 can be addressed by using a more efficient data structure for the stack.

6479 Exercises

- 6480 1. The dependency structure $1 \leftarrow 2 \rightarrow 3$, with 2 as the root, can be obtained from more
 6481 than one set of actions in arc-standard parsing. List both sets of actions that can
 6482 obtain this parse. Don't forget about the edge $\text{ROOT} \rightarrow 2$.
- 6483 2. This problem develops the relationship between dependency parsing and lexicalized
 6484 context-free parsing. Suppose you have a set of unlabeled arc scores $\{\psi(i \rightarrow j)\}_{i,j=1}^M \cup \{\psi(\text{ROOT} \rightarrow j)\}_{j=1}^M$.
 6485
 - 6486 a) Assuming each word type occurs no more than once in the input ($(i \neq j) \Rightarrow (w_i \neq w_j)$), how would you construct a weighted lexicalized context-free gram-
 6487 mar so that the score of *any* projective dependency tree is equal to the score of
 6488 some equivalent derivation in the lexicalized context-free grammar?
 - 6489 b) Verify that your method works for the example *They fish*.
 - 6490 c) Does your method require the restriction that each word type occur no more
 6491 than once in the input? If so, why?
 - 6493 d) *If your method required that each word type occur only once in the input,
 6494 show how to generalize it.

- 6495 3. In arc-factored dependency parsing of an input of length M , the score of a parse
 6496 is the sum of M scores, one for each arc. In second order dependency parsing, the
 6497 total score is the sum over many more terms. How many terms are the score of the
 6498 parse for Figure 11.2, using a second-order dependency parser with grandparent
 6499 and sibling features? Assume that a child of ROOT has no grandparent score, and
 6500 that a node with no siblings has no sibling scores.
- 6501 4. a) In the worst case, how many terms can be involved in the score of an input of
 6502 length M , assuming second-order dependency parsing? Describe the structure
 6503 of the worst-case parse. As in the previous problem, assume that there is only
 6504 one child of ROOT, and that it does not have any grandparent scores.
 6505 b) What about third-order dependency parsing?
- 6506 5. Provide the UD-style unlabeled dependency parse for the sentence *Xi-Lan eats shoots*
 6507 *and leaves*, assuming *shoots* is a noun and *leaves* is a verb. Provide arc-standard and
 6508 arc-eager derivations for this dependency parse.
- 6509 6. Compute an upper bound on the number of successful derivations in arc-standard
 6510 shift-reduce parsing for unlabeled dependencies, as a function of the length of the
 6511 input, M . Hint: a lower bound is the number of projective decision trees, $\frac{1}{M+1} \binom{3M-2}{M-1}$ (Zhang,
 6512 2017), where $\binom{a}{b} = \frac{a!}{(a-b)!b!}$.
- 6513 7. The **label bias problem** arises when a decision is locally correct, yet leads to a cas-
 6514 cade of errors in some situations (§ 11.3.3). Design a scenario in which this occurs.
 6515 Specifically:
- 6516 • Assume an arc-standard dependency parser, whose action classifier considers
 6517 only the words at the top of the stack and at the front of the input buffer.
 - 6518 • Design two examples, which both involve a decision with identical features.
 - 6519 – In one example, shift is the correct decision; in the other example, arc-left
 6520 or arc-right is the correct decision.
 - 6521 – In one of the two examples, a mistake should lead to at least two attach-
 6522 ment errors.
 - 6523 – In the other example, a mistake should lead only to a single attachment
 6524 error.

6525 For the following exercises, run a dependency parser, such as Stanford’s CoreNLP
 6526 parser, on a large corpus of text (at least 10^5 tokens), such as `nltk.corpus.webtext`.

- 6527 8. The dependency relation NMOD:POSS indicates possession. Compute the top ten
 6528 words most frequently possessed by each of the following pronouns: *his*, *her*, *our*,
 6529 *my*, *your*, and *their* (inspired by Muralidharan and Hearst, 2013).

9. Count all pairs of words grouped by the CONJ relation. Select all pairs of words (i, j) for which i and j each participate in CONJ relations at least five times. Compute and sort by the **pointwise mutual information**, which is defined in § 14.3 as,

$$\text{PMI}(i, j) = \log \frac{\text{p}(i, j)}{\text{p}(i)\text{p}(j)}. \quad [11.31]$$

Here, $\text{p}(i)$ is the fraction of CONJ relations containing word i (in either position), and $\text{p}(i, j)$ is the fraction of such relations linking i and j (in any order).

10. In § 4.2, we encountered lexical semantic relationships such as **synonymy** (same meaning), **antonymy** (opposite meaning), and **hyponymy** (i is a special case of j). Another relevant relation is **co-hyponymy**, which means that i and j share a hyponym. Of the top 20 pairs identified by PMI in the previous problem, how many participate in synsets that are linked by one of these four relations? Use WORDNET to check for these relations, and count a pair of words if any of their synsets are linked.

6539

Part III

6540

Meaning

6541 Chapter 12

6542 Logical semantics

6543 The previous few chapters have focused on building systems that reconstruct the **syntax**
6544 of natural language — its structural organization — through tagging and parsing. But
6545 some of the most exciting and promising potential applications of language technology
6546 involve going beyond syntax to **semantics** — the underlying meaning of the text:

- 6547 • Answering questions, such as *where is the nearest coffeeshop?* or *what is the middle name*
6548 *of the mother of the 44th President of the United States?*.
- 6549 • Building a robot that can follow natural language instructions to execute tasks.
- 6550 • Translating a sentence from one language into another, while preserving the under-
6551 lying meaning.
- 6552 • Fact-checking an article by searching the web for contradictory evidence.
- 6553 • Logic-checking an argument by identifying contradictions, ambiguity, and unsup-
6554 ported assertions.

6555 Semantic analysis involves converting natural language into a **meaning representa-**
6556 **tion**. To be useful, a meaning representation must meet several criteria:

- 6557 • **c1**: it should be unambiguous: unlike natural language, there should be exactly one
6558 meaning per statement;
- 6559 • **c2**: it should provide a way to link language to external knowledge, observations,
6560 and actions;
- 6561 • **c3**: it should support computational **inference**, so that meanings can be combined
6562 to derive additional knowledge;
- 6563 • **c4**: it should be expressive enough to cover the full range of things that people talk
6564 about in natural language.

6565 Much more than this can be said about the question of how best to represent knowledge
 6566 for computation (e.g., Sowa, 2000), but this chapter will focus on these four criteria.

6567 12.1 Meaning and denotation

6568 The first criterion for a meaning representation is that statements in the representation
 6569 should be unambiguous — they should have only one possible interpretation. Natural
 6570 language does not have this property: as we saw in chapter 10, sentences like *cats scratch*
 6571 *people with claws* have multiple interpretations.

6572 But what does it mean for a statement to be unambiguous? Programming languages
 6573 provide a useful example: the output of a program is completely specified by the rules of
 6574 the language and the properties of the environment in which the program is run. For ex-
 6575 ample, the python code $5 + 3$ will have the output 8, as will the codes $(4 * 4) - (3 * 3) + 1$
 6576 and $((8))$. This output is known as the **denotation** of the program, and can be written
 6577 as,

$$\llbracket 5+3 \rrbracket = \llbracket (4 * 4) - (3 * 3) + 1 \rrbracket = \llbracket ((8)) \rrbracket = 8. \quad [12.1]$$

6578 The denotations of these arithmetic expressions are determined by the meaning of the
 6579 **constants** (e.g., 5, 3) and the **relations** (e.g., $+$, $*$, $(,)$). Now let's consider another snippet
 6580 of python code, `double(4)`. The denotation of this code could be, $\llbracket \text{double}(4) \rrbracket = 8$, or
 6581 it could be $\llbracket \text{double}(4) \rrbracket = 44$ — it depends on the meaning of `double`. This meaning
 6582 is defined in a **world model** \mathcal{M} as an infinite set of pairs. We write the denotation with
 6583 respect to model \mathcal{M} as $\llbracket \cdot \rrbracket_{\mathcal{M}}$, e.g., $\llbracket \text{double} \rrbracket_{\mathcal{M}} = \{(0, 0), (1, 2), (2, 4), \dots\}$. The world
 6584 model would also define the (infinite) list of constants, e.g., $\{0, 1, 2, \dots\}$. As long as the
 6585 denotation of string ϕ in model \mathcal{M} can be computed unambiguously, the language can be
 6586 said to be unambiguous.

6587 This approach to meaning is known as **model-theoretic semantics**, and it addresses
 6588 not only criterion *c1* (no ambiguity), but also *c2* (connecting language to external knowl-
 6589 edge, observations, and actions). For example, we can connect a representation of the
 6590 meaning of a statement like *the capital of Georgia* with a world model that includes knowl-
 6591 edge base of geographical facts, obtaining the denotation `Atlanta`. We might populate a
 6592 world model by detecting and analyzing the objects in an image, and then use this world
 6593 model to evaluate **propositions** like *a man is riding a moose*. Another desirable property of
 6594 model-theoretic semantics is that when the facts change, the denotations change too: the
 6595 meaning representation of *President of the USA* would have a different denotation in the
 6596 model \mathcal{M}_{2014} as it would in \mathcal{M}_{2022} .

6597 12.2 Logical representations of meaning

6598 Criterion *c3* requires that the meaning representation support inference — for example,
 6599 automatically deducing new facts from known premises. While many representations
 6600 have been proposed that meet these criteria, the most mature is the language of first-order
 6601 logic.¹

6602 **12.2.1 Propositional logic**

6603 The bare bones of logical meaning representation are Boolean operations on propositions:

6604 **Propositional symbols.** Greek symbols like ϕ and ψ will be used to represent **proposi-**
 6605 **tions**, which are statements that are either true or false. For example, ϕ may corre-
 6606 spond to the proposition, *bagels are delicious*.

6607 **Boolean operators.** We can build up more complex propositional formulas from Boolean
 6608 operators. These include:

- 6609 • Negation $\neg\phi$, which is true if ϕ is false.
- 6610 • Conjunction, $\phi \wedge \psi$, which is true if both ϕ and ψ are true.
- 6611 • Disjunction, $\phi \vee \psi$, which is true if at least one of ϕ and ψ is true
- 6612 • Implication, $\phi \Rightarrow \psi$, which is true unless ϕ is true and ψ is false. Implication
 6613 has identical truth conditions to $\neg\phi \vee \psi$.
- 6614 • Equivalence, $\phi \Leftrightarrow \psi$, which is true if ϕ and ψ are both true or both false. Equiv-
 6615 alence has identical truth conditions to $(\phi \Rightarrow \psi) \wedge (\psi \Rightarrow \phi)$.

6616 It is not strictly necessary to have all five Boolean operators: readers familiar with
 6617 Boolean logic will know that it is possible to construct all other operators from either the
 6618 NAND (not-and) or NOR (not-or) operators. Nonetheless, it is clearest to use all five
 6619 operators. From the truth conditions for these operators, it is possible to define a number
 6620 of “laws” for these Boolean operators, such as,

- 6621 • *Commutativity*: $\phi \wedge \psi = \psi \wedge \phi$, $\phi \vee \psi = \psi \vee \phi$
- 6622 • *Associativity*: $\phi \wedge (\psi \wedge \chi) = (\phi \wedge \psi) \wedge \chi$, $\phi \vee (\psi \vee \chi) = (\phi \vee \psi) \vee \chi$
- 6623 • *Complementation*: $\phi \wedge \neg\phi = \perp$, $\phi \vee \neg\phi = \top$, where \top indicates a true proposition
 6624 and \perp indicates a false proposition.

¹Alternatives include the “variable-free” representation used in semantic parsing of geographical queries (Zelle and Mooney, 1996) and robotic control (Ge and Mooney, 2005), and dependency-based compositional semantics (Liang et al., 2013).

These laws can be combined to derive further equivalences, which can support logical inferences. For example, suppose $\phi = \text{The music is loud}$ and $\psi = \text{Max can't sleep}$. Then if we are given,

$$\begin{aligned}\phi \Rightarrow \psi & \quad \text{If the music is loud, Max can't sleep.} \\ \phi & \quad \text{The music is loud.}\end{aligned}$$

we can derive ψ (*Max can't sleep*) by application of **modus ponens**, which is one of a set of **inference rules** that can be derived from more basic laws and used to manipulate propositional formulas. **Automated theorem provers** are capable of applying inference rules to a set of premises to derive desired propositions (Loveland, 2016).

12.2.2 First-order logic

Propositional logic is so named because it treats propositions as its base units. However, the criterion *c4* states that our meaning representation should be sufficiently expressive. Now consider the sentence pair,

(12.1) If anyone is making noise, then Max can't sleep.
Abigail is making noise.

People are capable of making inferences from this sentence pair, but such inferences require formal tools that are beyond propositional logic. To understand the relationship between the statement *anyone is making noise* and the statement *Abigail is making noise*, our meaning representation requires the additional machinery of **first-order logic** (FOL).

In FOL, logical propositions can be constructed from relationships between entities. Specifically, FOL extends propositional logic with the following classes of terms:

Constants. These are elements that name individual entities in the model, such as MAX and ABIGAIL. The denotation of each constant in a model \mathcal{M} is an element in the model, e.g., $[\![\text{MAX}]\!] = m$ and $[\![\text{ABIGAIL}]\!] = a$.

Relations. Relations can be thought of as sets of entities, or sets of tuples. For example, the relation CAN-SLEEP is defined as the set of entities who can sleep, and has the denotation $[\![\text{CAN-SLEEP}]\!] = \{a, m, \dots\}$. To test the truth value of the proposition CAN-SLEEP(MAX), we ask whether $[\![\text{MAX}]\!] \in [\![\text{CAN-SLEEP}]\!]$. Logical relations that are defined over sets of entities are sometimes called *properties*.

Relations may also be ordered tuples of entities. For example BROTHER(MAX,ABIGAIL) expresses the proposition that MAX is the brother of ABIGAIL. The denotation of such relations is a set of tuples, $[\![\text{BROTHER}]\!] = \{(m, a), (x, y), \dots\}$. To test the truth value of the proposition BROTHER(MAX,ABIGAIL), we ask whether the tuple $([\![\text{MAX}]\!], [\![\text{ABIGAIL}]\!])$ is in the denotation $[\![\text{BROTHER}]\!]$.

Using constants and relations, it is possible to express statements like *Max can't sleep* and *Max is Abigail's brother*:

$$\neg\text{CAN-SLEEP}(\text{MAX}) \\ \text{BROTHER}(\text{MAX}, \text{ABIGAIL}).$$

These statements can also be combined using Boolean operators, such as,

$$(\text{BROTHER}(\text{MAX}, \text{ABIGAIL}) \vee \text{BROTHER}(\text{MAX}, \text{STEVE})) \Rightarrow \neg\text{CAN-SLEEP}(\text{MAX}).$$

6654 This fragment of first-order logic permits only statements about specific entities. To
 6655 support inferences about statements like *If anyone is making noise, then Max can't sleep*,
 6656 two more elements must be added to the meaning representation:

6657 **Variables.** Variables are mechanisms for referring to entities that are not locally specified.
 6658 We can then write $\text{CAN-SLEEP}(x)$ or $\text{BROTHER}(x, \text{ABIGAIL})$. In these cases, x is a **free**
 6659 **variable**, meaning that we have not committed to any particular assignment.

6660 **Quantifiers.** Variables are bound by quantifiers. There are two quantifiers in first-order
 6661 logic.²

- 6662 • The **existential quantifier** \exists , which indicates that there must be at least one en-
 6663 tity to which the variable can bind. For example, the statement $\exists x \text{MAKES-NOISE}(x)$
 6664 indicates that there is at least one entity for which MAKES-NOISE is true.
- 6665 • The **universal quantifier** \forall , which indicates that the variable must be able to
 6666 bind to any entity in the model. For example, the statement,

$$\text{MAKES-NOISE}(\text{ABIGAIL}) \Rightarrow (\forall x \neg \text{CAN-SLEEP}(x)) \quad [12.3]$$

6667 asserts that if Abigail makes noise, no one can sleep.

6668 The expressions $\exists x$ and $\forall x$ make x into a **bound variable**. A formula that contains
 6669 no free variables is a **sentence**.

6670 **Functions.** Functions map from entities to entities, e.g., $\llbracket \text{CAPITAL-OF(GEORGIA)} \rrbracket = \llbracket \text{ATLANTA} \rrbracket$.
 6671 With functions, it is convenient to add an equality operator, supporting statements
 6672 like,

$$\forall x \exists y \text{MOTHER-OF}(x) = \text{DAUGHTER-OF}(y). \quad [12.4]$$

²In first-order logic, it is possible to quantify only over entities. In **second-order logic**, it is possible to quantify over properties. This makes it possible to represent statements like *Butch has every property that a good boxer has* (example from Blackburn and Bos, 2005),

$$\forall P \forall x ((\text{GOOD-BOXER}(x) \Rightarrow P(x)) \Rightarrow P(\text{BUTCH})). \quad [12.2]$$

6673 Note that MOTHER-OF is a functional analogue of the relation MOTHER, so that
 6674 $\text{MOTHER-OF}(x) = y$ if $\text{MOTHER}(x, y)$. Any logical formula that uses functions can be
 6675 rewritten using only relations and quantification. For example,

$$\text{MAKES-NOISE}(\text{MOTHER-OF}(\text{ABIGAIL})) \quad [12.5]$$

6676 can be rewritten as $\exists x \text{MAKES-NOISE}(x) \wedge \text{MOTHER}(x, \text{ABIGAIL})$.

An important property of quantifiers is that the order can matter. Unfortunately, natural language is rarely clear about this! The issue is demonstrated by examples like *everyone speaks a language*, which has the following interpretations:

$$\forall x \exists y \text{ SPEAKS}(x, y) \quad [12.6]$$

$$\exists y \forall x \text{ SPEAKS}(x, y). \quad [12.7]$$

6677 In the first case, y may refer to several different languages, while in the second case, there
 6678 is a single y that is spoken by everyone.

6679 Truth-conditional semantics

6680 One way to look at the meaning of an FOL sentence ϕ is as a set of **truth conditions**,
 6681 or models under which ϕ is satisfied. But how to determine whether a sentence is true
 6682 or false in a given model? We will approach this inductively, starting with a predicate
 6683 applied to a tuple of constants. The truth of such a sentence depends on whether the
 6684 tuple of denotations of the constants is in the denotation of the predicate. For example,
 6685 $\text{CAPITAL}(\text{GEORGIA}, \text{ATLANTA})$ is true in model \mathcal{M} iff,

$$(\llbracket \text{GEORGIA} \rrbracket_{\mathcal{M}}, \llbracket \text{ATLANTA} \rrbracket_{\mathcal{M}}) \in \llbracket \text{CAPITAL} \rrbracket_{\mathcal{M}}. \quad [12.8]$$

6686 The Boolean operators \wedge, \vee, \dots provide ways to construct more complicated sentences,
 6687 and the truth of such statements can be assessed based on the truth tables associated with
 6688 these operators. The statement $\exists x \phi$ is true if there is some assignment of the variable x
 6689 to an entity in the model such that ϕ is true; the statement $\forall x \phi$ is true if ϕ is true under
 6690 all possible assignments of x . More formally, we would say that ϕ is **satisfied** under \mathcal{M} ,
 6691 written as $\mathcal{M} \models \phi$.

6692 Truth conditional semantics allows us to define several other properties of sentences
 6693 and pairs of sentences. Suppose that in every \mathcal{M} under which ϕ is satisfied, another
 6694 formula ψ is also satisfied; then ϕ **entails** ψ , which is also written as $\phi \models \psi$. For example,

$$\text{CAPITAL}(\text{GEORGIA}, \text{ATLANTA}) \models \exists x \text{CAPITAL}(\text{GEORGIA}, x). \quad [12.9]$$

6695 A statement that is satisfied under any model, such as $\phi \vee \neg\phi$, is **valid**, written $\models (\phi \vee$
 6696 $\neg\phi)$. A statement that is not satisfied under any model, such as $\phi \wedge \neg\phi$, is **unsatisfiable**,

6697 or **inconsistent**. A **model checker** is a program that determines whether a sentence ϕ
6698 is satisfied in \mathcal{M} . A **model builder** is a program that constructs a model in which ϕ
6699 is satisfied. The problems of checking for consistency and validity in first-order logic
6700 are **undecidable**, meaning that there is no algorithm that can automatically determine
6701 whether an FOL formula is valid or inconsistent.

6702 **Inference in first-order logic**

6703 Our original goal was to support inferences that combine general statements *If anyone is*
6704 *making noise, then Max can't sleep* with specific statements like *Abigail is making noise*. We
6705 can now represent such statements in first-order logic, but how are we to perform the
6706 inference that *Max can't sleep*? One approach is to use “generalized” versions of proposi-
6707 tional inference rules like modus ponens, which can be applied to FOL formulas. By
6708 repeatedly applying such inference rules to a knowledge base of facts, it is possible to
6709 produce proofs of desired propositions. To find the right sequence of inferences to derive
6710 a desired theorem, classical artificial intelligence search algorithms like backward chain-
6711 ing can be applied. Such algorithms are implemented in interpreters for the `prolog` logic
6712 programming language (Pereira and Shieber, 2002).

6713 **12.3 Semantic parsing and the lambda calculus**

6714 The previous section laid out a lot of formal machinery; the remainder of this chapter
6715 links these formalisms back to natural language. Given an English sentence like *Alex likes*
6716 *Brit*, how can we obtain the desired first-order logical representation, `LIKES(ALEX,BRIT)`?
6717 This is the task of **semantic parsing**. Just as a syntactic parser is a function from a natu-
6718 ral language sentence to a syntactic structure such as a phrase structure tree, a semantic
6719 parser is a function from natural language to logical formulas.

6720 As in syntactic analysis, semantic parsing is difficult because the space of inputs and
6721 outputs is very large, and their interaction is complex. Our best hope is that, like syntactic
6722 parsing, semantic parsing can somehow be decomposed into simpler sub-problems. This
6723 idea, usually attributed to the German philosopher Gottlob Frege, is called the **principle**
6724 **of compositionality**: the meaning of a complex expression is a function of the meanings of
6725 that expression’s constituent parts. We will define these “constituent parts” as syntactic
6726 constituents: noun phrases and verb phrases. These constituents are combined using
6727 function application: if the syntactic parse contains the production $x \rightarrow y z$, then the
6728 semantics of x , written $x.\text{sem}$, will be computed as a function of the semantics of the

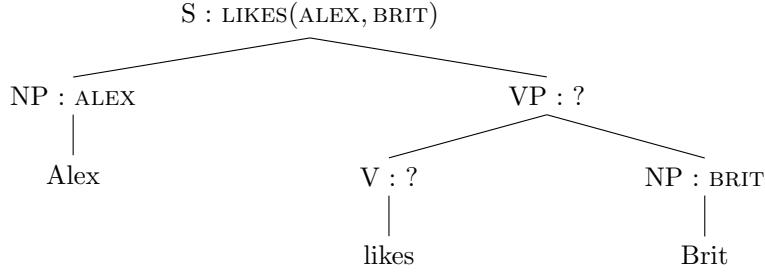


Figure 12.1: The principle of compositionality requires that we identify meanings for the constituents *likes* and *likes Brit* that will make it possible to compute the meaning for the entire sentence.

6729 constituents, $y.\text{sem}$ and $z.\text{sem}$.³ ⁴

6730 12.3.1 The lambda calculus

6731 Let's see how this works for a simple sentence like *Alex likes Brit*, whose syntactic structure
 6732 is shown in Figure 12.1. Our goal is the formula, $\text{LIKES}(\text{ALEX}, \text{BRIT})$, and it is clear that the
 6733 meaning of the constituents *Alex* and *Brit* should be *ALEX* and *BRIT*. That leaves two more
 6734 constituents: the verb *likes*, and the verb phrase *likes Brit*. The meanings of these units
 6735 must be defined in a way that makes it possible to recover the desired meaning for the
 6736 entire sentence by function application. If the meanings of *Alex* and *Brit* are constants,
 6737 then the meanings of *likes* and *likes Brit* must be functional expressions, which can be
 6738 applied to their siblings to produce the desired analyses.

6739 Modeling these partial analyses requires extending the first-order logic meaning rep-
 6740 resentation. We do this by adding **lambda expressions**, which are descriptions of anony-
 6741 mous functions,⁵ e.g.,

$$\lambda x. \text{LIKES}(x, \text{BRIT}). \quad [12.10]$$

6742 This functional expression is the meaning of the verb phrase *likes Brit*; it takes a single
 6743 argument, and returns the result of substituting that argument for x in the expression

³§ 9.3.2 briefly discusses Combinatory Categorial Grammar (CCG) as an alternative to a phrase-structure analysis of syntax. CCG is argued to be particularly well-suited to semantic parsing (Hockenmaier and Steedman, 2007), and is used in much of the contemporary work on machine learning for semantic parsing, summarized in § 12.4.

⁴The approach of algorithmically building up meaning representations from a series of operations on the syntactic structure of a sentence is generally attributed to the philosopher Richard Montague, who published a series of influential papers on the topic in the early 1970s (e.g., Montague, 1973).

⁵Formally, all first-order logic formulas are lambda expressions; in addition, if ϕ is a lambda expression, then $\lambda x.\phi$ is also a lambda expression. Readers who are familiar with functional programming will recognize lambda expressions from their use in programming languages such as Lisp and Python.

6744 $\text{LIKES}(x, \text{BRIT})$. We write this substitution as,

$$(\lambda x.\text{LIKES}(x, \text{BRIT}))@\text{ALEX} = \text{LIKES}(\text{ALEX}, \text{BRIT}), \quad [12.11]$$

6745 with the symbol “@” indicating function application. Function application in the lambda
 6746 calculus is sometimes called **β -reduction** or β -conversion. The expression $\phi@\psi$ indicates
 6747 a function application to be performed by β -reduction, and $\phi(\psi)$ indicates a function or
 6748 predicate in the final logical form.

6749 Equation 12.11 shows how to obtain the desired semantics for the sentence *Alex likes*
 6750 *Brit*: by applying the lambda expression $\lambda x.\text{LIKES}(x, \text{BRIT})$ to the logical constant ALEX.
 6751 This rule of composition can be specified in a **syntactic-semantic grammar**, in which
 6752 syntactic productions are paired with semantic operations. For the syntactic production
 6753 $S \rightarrow \text{NP VP}$, we have the semantic rule $\text{VP.sem}@\text{NP.sem}$.

The meaning of the transitive verb phrase *likes Brit* can also be obtained by function application on its syntactic constituents. For the syntactic production $\text{VP} \rightarrow \text{V NP}$, we apply the semantic rule,

$$\text{VP.sem} = (\text{V.sem})@\text{NP.sem} \quad [12.12]$$

$$= (\lambda y.\lambda x.\text{LIKES}(x, y)) @ (\text{BRIT}) \quad [12.13]$$

$$= \lambda x.\text{LIKES}(x, \text{BRIT}). \quad [12.14]$$

6754 Thus, the meaning of the transitive verb *likes* is a lambda expression whose output is
 6755 *another* lambda expression: it takes y as an argument to fill in one of the slots in the LIKES
 6756 relation, and returns a lambda expression that is ready to take an argument to fill in the
 6757 other slot.⁶

6758 Table 12.1 shows a minimal syntactic-semantic grammar fragment, G_1 . The complete
 6759 **derivation** of *Alex likes Brit* in G_1 is shown in Figure 12.2. In addition to the transitive
 6760 verb *likes*, the grammar also includes the intransitive verb *sleeps*; it should be clear how
 6761 to derive the meaning of sentences like *Alex sleeps*. For verbs that can be either transitive
 6762 or intransitive, such as *eats*, we would have two terminal productions, one for each sense
 6763 (terminal productions are also called the **lexical entries**). Indeed, most of the grammar is
 6764 in the **lexicon** (the terminal productions), since these productions select the basic units of
 6765 the semantic interpretation.

6766 12.3.2 Quantification

6767 Things get more complicated when we move from sentences about named entities to sen-
 6768 tences that involve more general noun phrases. Let’s consider the example, *A dog sleeps*,

⁶This can be written in a few different ways. The notation $\lambda y, x.\text{LIKES}(x, y)$ is a somewhat informal way to indicate a lambda expression that takes two arguments; this would be acceptable in functional programming. Logicians (e.g., Carpenter, 1997) often prefer the more formal notation $\lambda y.\lambda x.\text{LIKES}(x)(y)$, indicating that each lambda expression takes exactly one argument.

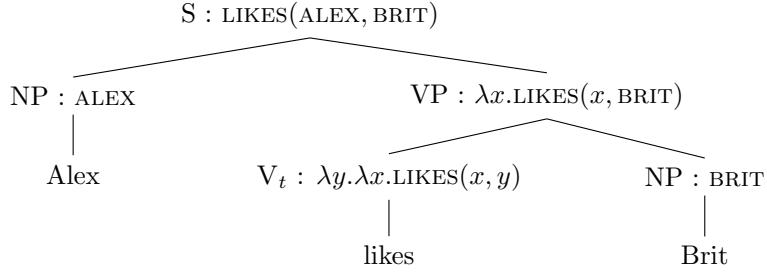


Figure 12.2: Derivation of the semantic representation for *Alex likes Brit* in the grammar G_1 .

S	\rightarrow	NP VP	VP.sem@NP.sem
VP	\rightarrow	V _t NP	V _t .sem@NP.sem
VP	\rightarrow	V _i	V _i .sem
V _t	\rightarrow	likes	$\lambda y. \lambda x. \text{LIKES}(x, y)$
V _i	\rightarrow	sleeps	$\lambda x. \text{SLEEPS}(x)$
NP	\rightarrow	Alex	ALEX
NP	\rightarrow	Brit	BRIT

Table 12.1: G_1 , a minimal syntactic-semantic context-free grammar

which has the meaning $\exists x \text{DOG}(x) \wedge \text{SLEEPS}(x)$. Clearly, the DOG relation will be introduced by the word *dog*, and the SLEEP relation will be introduced by the word *sleeps*. The existential quantifier \exists must be introduced by the lexical entry for the determiner *a*.⁷ However, this seems problematic for the compositional approach taken in the grammar G_1 : if the semantics of the noun phrase *a dog* is an existentially quantified expression, how can it be the argument to the semantics of the verb *sleeps*, which expects an entity? And where does the logical conjunction come from?

There are a few different approaches to handling these issues.⁸ We will begin by reversing the semantic relationship between subject NPs and VPs, so that the production $S \rightarrow \text{NP VP}$ has the semantics $\text{NP.sem}@\text{VP.sem}$: the meaning of the sentence is now the semantics of the noun phrase applied to the verb phrase. The implications of this change are best illustrated by exploring the derivation of the example, shown in Figure 12.3. Let's

⁷Conversely, the sentence *Every dog sleeps* would involve a universal quantifier, $\forall x \text{DOG}(x) \Rightarrow \text{SLEEPS}(x)$. The definite article *the* requires more consideration, since *the dog* must refer to some dog which is uniquely identifiable, perhaps from contextual information external to the sentence. Carpenter (1997, pp. 96-100) summarizes recent approaches to handling definite descriptions.

⁸Carpenter (1997) offers an alternative treatment based on combinatory categorial grammar.

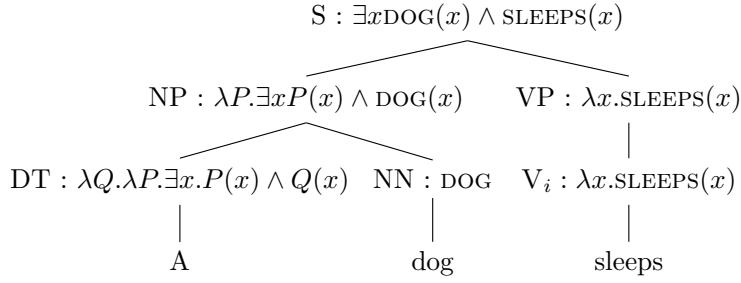


Figure 12.3: Derivation of the semantic representation for *A dog sleeps*, in grammar G_2

6781 start with the indefinite article *a*, to which we assign the rather intimidating semantics,

$$\lambda P. \lambda Q. \exists x P(x) \wedge Q(x). \quad [12.15]$$

This is a lambda expression that takes two **relations** as arguments, P and Q . The relation P is scoped to the outer lambda expression, so it will be provided by the immediately adjacent noun, which in this case is DOG. Thus, the noun phrase *a dog* has the semantics,

$$\text{NP.sem} = \text{DET.sem} @ \text{NN.sem} \quad [12.16]$$

$$= (\lambda P. \lambda Q. \exists x P(x) \wedge Q(x)) @ (\text{DOG}) \quad [12.17]$$

$$= \lambda Q. \exists x \text{DOG}(x) \wedge Q(x). \quad [12.18]$$

6782 This is a lambda expression that is expecting another relation, Q , which will be provided
6783 by the verb phrase, SLEEPS. This gives the desired analysis, $\exists x \text{DOG}(x) \wedge \text{SLEEPS}(x)$.⁹

6784 If noun phrases like *a dog* are interpreted as lambda expressions, then proper nouns
6785 like *Alex* must be treated in the same way. This is achieved by **type-raising** from con-
6786 stants to lambda expressions, $x \Rightarrow \lambda P. P(x)$. After type-raising, the semantics of *Alex* is
6787 $\lambda P. P(\text{ALEX})$ — a lambda expression that expects a relation to tell us something about
6788 *ALEX*.¹⁰ Again, make sure you see how the analysis in Figure 12.3 can be applied to the
6789 sentence *Alex sleeps*.

⁹When applying β -reduction to arguments that are themselves lambda expressions, be sure to use unique variable names to avoid confusion. For example, it is important to distinguish the x in the semantics for *a* from the x in the semantics for *likes*. Variable names are abstractions, and can always be changed — this is known as **α -conversion**. For example, $\lambda x. P(x)$ can be converted to $\lambda y. P(y)$, etc.

¹⁰Compositional semantic analysis is often supported by **type systems**, which make it possible to check whether a given function application is valid. The base types are entities e and truth values t . A property, such as DOG, is a function from entities to truth values, so its type is written $\langle e, t \rangle$. A transitive verb has type $\langle e, \langle e, t \rangle \rangle$: after receiving the first entity (the direct object), it returns a function from entities to truth values, which will be applied to the subject of the sentence. The type-raising operation $x \Rightarrow \lambda P. P(x)$ corresponds to a change in type from e to $\langle \langle e, t \rangle, t \rangle$: it expects a function from entities to truth values, and returns a truth value.

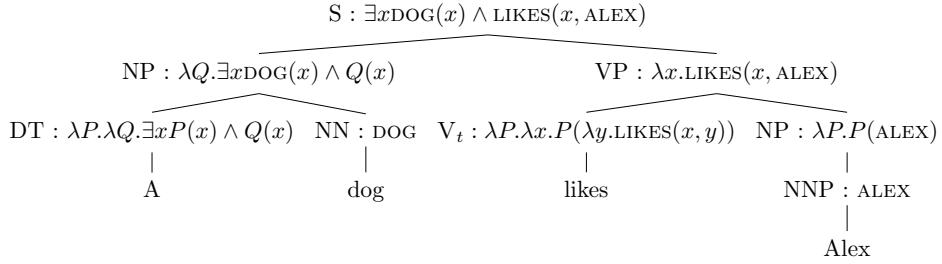


Figure 12.4: Derivation of the semantic representation for *A dog likes Alex*.

6790 Direct objects are handled by applying the same type-raising operation to transitive
 6791 verbs: the meaning of verbs such as *likes* is raised to,

$$\lambda P. \lambda x. P(\lambda y. \text{LIKES}(x, y)) \quad [12.19]$$

As a result, we can keep the verb phrase production $\text{VP.sem} = \text{V.sem}@\text{NP.sem}$, knowing that the direct object will provide the function P in Equation 12.19. To see how this works, let's analyze the verb phrase *likes a dog*. After uniquely relabeling each lambda variable,

$$\begin{aligned}
 \text{VP.sem} &= \text{V.sem}@\text{NP.sem} \\
 &= (\lambda P. \lambda x. P(\lambda y. \text{LIKES}(x, y))) @ (\lambda Q. \exists z \text{DOG}(z) \wedge Q(z)) \\
 &= \lambda x. (\lambda Q. \exists z \text{DOG}(z) \wedge Q(z)) @ (\lambda y. \text{LIKES}(x, y)) \\
 &= \lambda x. \exists z \text{DOG}(z) \wedge (\lambda y. \text{LIKES}(x, y)) @ z \\
 &= \lambda x. \exists z \text{DOG}(z) \wedge \text{LIKES}(x, z).
 \end{aligned}$$

6792 These changes are summarized in the revised grammar G_2 , shown in Table 12.2. Figure 12.4 shows a derivation that involves a transitive verb, an indefinite noun phrase, and
 6793 a proper noun.

6795 12.4 Learning semantic parsers

6796 As with syntactic parsing, any syntactic-semantic grammar with sufficient coverage risks
 6797 producing many possible analyses for any given sentence. Machine learning is the dom-
 6798 inant approach to selecting a single analysis. We will focus on algorithms that learn to
 6799 score logical forms by attaching weights to features of their derivations (Zettlemoyer
 6800 and Collins, 2005). Alternative approaches include transition-based parsing (Zelle and
 6801 Mooney, 1996; Misra and Artzi, 2016) and methods inspired by machine translation (Wong
 6802 and Mooney, 2006). Methods also differ in the form of supervision used for learning,
 6803 which can range from complete derivations to much more limited training signals. We
 6804 will begin with the case of complete supervision, and then consider how learning is still
 6805 possible even when seemingly key information is missing.

S	\rightarrow NP VP	NP.sem@VP.sem
VP	\rightarrow V_t NP	$V_t.sem@NP.sem$
VP	\rightarrow V_i	$V_i.sem$
NP	\rightarrow DET NN	DET.sem@NN.sem
NP	\rightarrow NNP	$\lambda P.P(NNP.sem)$
DET	$\rightarrow a$	$\lambda P.\lambda Q.\exists xP(x) \wedge Q(x)$
DET	$\rightarrow every$	$\lambda P.\lambda Q.\forall x(P(x) \Rightarrow Q(x))$
V_t	$\rightarrow likes$	$\lambda P.\lambda x.P(\lambda y.LIKES(x,y))$
V_i	$\rightarrow sleeps$	$\lambda x.SLEEPS(x)$
NN	$\rightarrow dog$	DOG
NNP	$\rightarrow Alex$	ALEX
NNP	$\rightarrow Brit$	BRIT

Table 12.2: G_2 , a syntactic-semantic context-free grammar fragment, which supports quantified noun phrases

6806 **Datasets** Early work on semantic parsing focused on natural language expressions of
 6807 geographical database queries, such as *What states border Texas*. The GeoQuery dataset
 6808 of Zelle and Mooney (1996) was originally coded in prolog, but has subsequently been
 6809 expanded and converted into the SQL database query language by Popescu et al. (2003)
 6810 and into first-order logic with lambda calculus by Zettlemoyer and Collins (2005), pro-
 6811 viding logical forms like $\lambda x.\text{STATE}(x) \wedge \text{BORDERS}(x, \text{TEXAS})$. Another early dataset con-
 6812 sists of instructions for RoboCup robot soccer teams (Kate et al., 2005). More recent work
 6813 has focused on broader domains, such as the Freebase database (Bollacker et al., 2008),
 6814 for which queries have been annotated by Krishnamurthy and Mitchell (2012) and Cai
 6815 and Yates (2013). Other recent datasets include child-directed speech (Kwiatkowski et al.,
 6816 2012) and elementary school science exams (Krishnamurthy, 2016).

6817 12.4.1 Learning from derivations

Let $w^{(i)}$ indicate a sequence of text, and let $y^{(i)}$ indicate the desired logical form. For example:

$$\begin{aligned} w^{(i)} &= \text{Alex eats shoots and leaves} \\ y^{(i)} &= \text{EATS(ALEX,SHOOTS)} \wedge \text{EATS(ALEX,LEAVES)} \end{aligned}$$

6818 In the standard supervised learning paradigm that was introduced in § 2.3, we first de-
 6819 fine a feature function, $f(w, y)$, and then learn weights on these features, so that $y^{(i)} =$
 6820 $\text{argmax}_y \theta \cdot f(w, y)$. The weight vector θ is learned by comparing the features of the true
 6821 label $f(w^{(i)}, y^{(i)})$ against either the features of the predicted label $f(w^{(i)}, \hat{y})$ (perceptron,

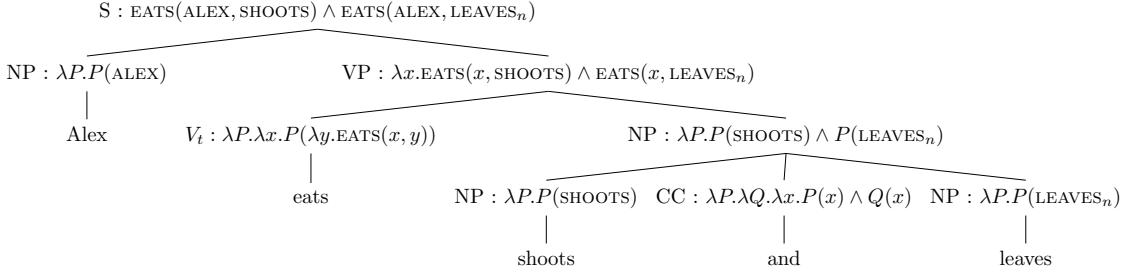


Figure 12.5: Derivation for gold semantic analysis of *Alex eats shoots and leaves*

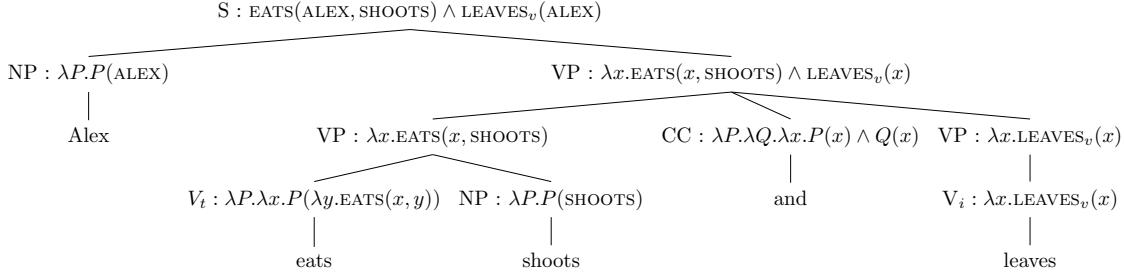
support vector machine) or the expected feature vector $E_{\mathbf{y}|\mathbf{w}}[\mathbf{f}(\mathbf{w}^{(i)}, \mathbf{y})]$ (logistic regression).

While this basic framework seems similar to discriminative syntactic parsing, there is a crucial difference. In (context-free) syntactic parsing, the annotation $\mathbf{y}^{(i)}$ contains all of the syntactic productions; indeed, the task of identifying the correct set of productions is identical to the task of identifying the syntactic structure. In semantic parsing, this is not the case: the logical form $EATS(ALEX, SHOOTS) \wedge EATS(ALEX, LEAVES)$ does not reveal the syntactic-semantic productions that were used to obtain it. Indeed, there may be **spurious ambiguity**, so that a single logical form can be reached by multiple derivations. (We previously encountered spurious ambiguity in transition-based dependency parsing, § 11.3.2.)

These ideas can be formalized by introducing an additional variable \mathbf{z} , representing the derivation of the logical form \mathbf{y} from the text \mathbf{w} . Assume that the feature function decomposes across the productions in the derivation, $\mathbf{f}(\mathbf{w}, \mathbf{z}, \mathbf{y}) = \sum_{t=1}^T \mathbf{f}(\mathbf{w}, z_t, \mathbf{y})$, where z_t indicates a single syntactic-semantic production. For example, we might have a feature for the production $S \rightarrow NP VP : NP.sem@VP.sem$, as well as for terminal productions like $NNP \rightarrow Alex : ALEX$. Under this decomposition, it is possible to compute scores for each semantically-annotated subtree in the analysis of \mathbf{w} , so that bottom-up parsing algorithms like CKY (§ 10.1) can be applied to find the best-scoring semantic analysis.

Figure 12.5 shows a derivation of the correct semantic analysis of the sentence *Alex eats shoots and leaves*, in a simplified grammar in which the plural noun phrases *shoots* and *leaves* are interpreted as logical constants *SHOOTS* and *LEAVES_n*. Figure 12.6 shows a derivation of an incorrect analysis. Assuming one feature per production, the perceptron update is shown in Table 12.3. From this update, the parser would learn to prefer the noun interpretation of *leaves* over the verb interpretation. It would also learn to prefer noun phrase coordination over verb phrase coordination.

While the update is explained in terms of the perceptron, it would be easy to replace the perceptron with a conditional random field. In this case, the online updates would be

Figure 12.6: Derivation for incorrect semantic analysis of *Alex eats shoots and leaves*

$NP_1 \rightarrow NP_2 \text{ CC } NP_3$	$(CC.\text{sem} @ (NP_2.\text{sem})) @ (NP_3.\text{sem})$	+1
$VP_1 \rightarrow VP_2 \text{ CC } VP_3$	$(CC.\text{sem} @ (VP_2.\text{sem})) @ (VP_3.\text{sem})$	-1
$NP \rightarrow leaves$	LEAVES_n	+1
$VP \rightarrow V_i$	$V_i.\text{sem}$	-1
$V_i \rightarrow leaves$	$\lambda x.\text{LEAVES}_v$	-1

Table 12.3: Perceptron update for analysis in Figure 12.5 (gold) and Figure 12.6 (predicted)

6850 based on feature expectations, which can be computed using the inside-outside algorithm
 6851 (§ 10.6).

6852 12.4.2 Learning from logical forms

Complete derivations are expensive to annotate, and are rarely available.¹¹ One solution is to focus on learning from logical forms directly, while treating the derivations as **latent variables** (Zettlemoyer and Collins, 2005). In a conditional probabilistic model over logical forms y and derivations z , we have,

$$p(y, z | w) = \frac{\exp(\theta \cdot f(w, z, y))}{\sum_{y', z'} \exp(\theta \cdot f(w, z', y'))}, \quad [12.20]$$

6853 which is the standard log-linear model, applied to the logical form y and the derivation
 6854 z .

Since the derivation z unambiguously determines the logical form y , it may seem silly to model the joint probability over y and z . However, since z is unknown, it can be marginalized out,

$$p(y | w) = \sum_z p(y, z | w). \quad [12.21]$$

¹¹An exception is the work of Ge and Mooney (2005), who annotate the meaning of each syntactic constituents for several hundred sentences.

The semantic parser can then select the logical form with the maximum log marginal probability,

$$\log \sum_z p(\mathbf{y}, \mathbf{z} \mid \mathbf{w}) = \log \sum_z \frac{\exp(\boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{w}, \mathbf{z}, \mathbf{y}))}{\sum_{\mathbf{y}', \mathbf{z}' \exp(\boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{w}, \mathbf{z}', \mathbf{y}'))}} \quad [12.22]$$

$$\propto \log \sum_z \exp(\boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{w}, \mathbf{z}', \mathbf{y}')) \quad [12.23]$$

$$\geq \max_z \boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{w}, \mathbf{z}, \mathbf{y}). \quad [12.24]$$

It is impossible to push the log term inside the sum over \mathbf{z} , so our usual linear scoring function does not apply. We can recover this scoring function only in approximation, by taking the max (rather than the sum) over derivations \mathbf{z} , which provides a lower bound.

Learning can be performed by maximizing the log marginal likelihood,

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^N \log p(\mathbf{y}^{(i)} \mid \mathbf{w}^{(i)}; \boldsymbol{\theta}) \quad [12.25]$$

$$= \sum_{i=1}^N \log \sum_{\mathbf{z}} p(\mathbf{y}^{(i)}, \mathbf{z}^{(i)} \mid \mathbf{w}^{(i)}; \boldsymbol{\theta}). \quad [12.26]$$

This log-likelihood is not **convex** in $\boldsymbol{\theta}$, unlike the log-likelihood of a fully-observed conditional random field. This means that learning can give different results depending on the initialization.

The derivative of Equation 12.26 is,

$$\frac{\partial \ell_i}{\partial \boldsymbol{\theta}} = \sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{y}, \mathbf{w}; \boldsymbol{\theta}) \mathbf{f}(\mathbf{w}, \mathbf{z}, \mathbf{y}) - \sum_{\mathbf{y}', \mathbf{z}'} p(\mathbf{y}', \mathbf{z}' \mid \mathbf{w}; \boldsymbol{\theta}) \mathbf{f}(\mathbf{w}, \mathbf{z}', \mathbf{y}') \quad [12.27]$$

$$= E_{\mathbf{z} \mid \mathbf{y}, \mathbf{w}} \mathbf{f}(\mathbf{w}, \mathbf{z}, \mathbf{y}) - E_{\mathbf{y}, \mathbf{z} \mid \mathbf{w}} \mathbf{f}(\mathbf{w}, \mathbf{z}, \mathbf{y}) \quad [12.28]$$

Both expectations can be computed via bottom-up algorithms like inside-outside. Alternatively, we can again maximize rather than marginalize over derivations for an approximate solution. In either case, the first term of the gradient requires us to identify derivations \mathbf{z} that are compatible with the logical form \mathbf{y} . This can be done in a bottom-up dynamic programming algorithm, by having each cell in the table $t[i, j, X]$ include the set of all possible logical forms for $X \sim \mathbf{w}_{i+1:j}$. The resulting table may therefore be much larger than in syntactic parsing. This can be controlled by using pruning to eliminate intermediate analyses that are incompatible with the final logical form \mathbf{y} (Zettlemoyer and Collins, 2005), or by using beam search and restricting the size of each cell to some fixed constant (Liang et al., 2013).

If we replace each expectation in Equation 12.28 with argmax and then apply stochastic gradient descent to learn the weights, we obtain the **latent variable perceptron**, a simple

Algorithm 16 Latent variable perceptron

```

1: procedure LATENTVARIABLEPERCEPTRON( $w^{(1:N)}, y^{(1:N)}$ )
2:    $\theta \leftarrow 0$ 
3:   repeat
4:     Select an instance  $i$ 
5:      $z^{(i)} \leftarrow \text{argmax}_z \theta \cdot f(w^{(i)}, z, y^{(i)})$ 
6:      $\hat{y}, \hat{z} \leftarrow \text{argmax}_{y', z'} \theta \cdot f(w^{(i)}, z', y')$ 
7:      $\theta \leftarrow \theta + f(w^{(i)}, z^{(i)}, y^{(i)}) - f(w^{(i)}, \hat{z}, \hat{y})$ 
8:   until tired
9:   return  $\theta$ 

```

and general algorithm for learning with missing data. The algorithm is shown in its most basic form in Algorithm 16, but the usual tricks such as averaging and margin loss can be applied (Yu and Joachims, 2009). Aside from semantic parsing, the latent variable perceptron has been used in tasks such as machine translation (Liang et al., 2006) and named entity recognition (Sun et al., 2009). In **latent conditional random fields**, we use the full expectations rather than maximizing over the hidden variable. This model has also been employed in a range of problems beyond semantic parsing, including parse reranking (Koo and Collins, 2005) and gesture recognition (Quattoni et al., 2007).

12.4.3 Learning from denotations

Logical forms are easier to obtain than complete derivations, but the annotation of logical forms still requires considerable expertise. However, it is relatively easy to obtain denotations for many natural language sentences. For example, in the geography domain, the denotation of a question would be its answer (Clarke et al., 2010; Liang et al., 2013):

Text :What states border Georgia?
Logical form : $\lambda x.\text{STATE}(x) \wedge \text{BORDER}(x, \text{GEORGIA})$
Denotation :{Alabama, Florida, North Carolina,
 South Carolina, Tennessee}

Similarly, in a robotic control setting, the denotation of a command would be an action or sequence of actions (Artzi and Zettlemoyer, 2013). In both cases, the idea is to reward the semantic parser for choosing an analysis whose denotation is correct: the right answer to the question, or the right action.

Learning from logical forms was made possible by summing or maxing over derivations. This idea can be carried one step further, summing or maxing over all logical forms with the correct denotation. Let $v_i(y) \in \{0, 1\}$ be a **validation function**, which assigns a

binary score indicating whether the denotation $\llbracket \mathbf{y} \rrbracket$ for the text $\mathbf{w}^{(i)}$ is correct. We can then learn by maximizing a conditional-likelihood objective,

$$\ell^{(i)}(\boldsymbol{\theta}) = \log \sum_{\mathbf{y}} v_i(\mathbf{y}) \times p(\mathbf{y} \mid \mathbf{w}; \boldsymbol{\theta}) \quad [12.29]$$

$$= \log \sum_{\mathbf{y}} v_i(\mathbf{y}) \times \sum_z p(\mathbf{y}, z \mid \mathbf{w}; \boldsymbol{\theta}), \quad [12.30]$$

which sums over all derivations z of all valid logical forms, $\{\mathbf{y} : v_i(\mathbf{y}) = 1\}$. This corresponds to the log-probability that the semantic parser produces a logical form with a valid denotation.

Differentiating with respect to $\boldsymbol{\theta}$, we obtain,

$$\frac{\partial \ell^{(i)}}{\partial \boldsymbol{\theta}} = \sum_{\mathbf{y}, z: v_i(\mathbf{y})=1} p(\mathbf{y}, z \mid \mathbf{w}) f(\mathbf{w}, z, \mathbf{y}) - \sum_{\mathbf{y}', z'} p(\mathbf{y}', z' \mid \mathbf{w}) f(\mathbf{w}, z', \mathbf{y}'), \quad [12.31]$$

which is the usual difference in feature expectations. The positive term computes the expected feature expectations conditioned on the denotation being valid, while the second term computes the expected feature expectations according to the current model, without regard to the ground truth. Large-margin learning formulations are also possible for this problem. For example, Artzi and Zettlemoyer (2013) generate a set of valid and invalid derivations, and then impose a constraint that all valid derivations should score higher than all invalid derivations. This constraint drives a perceptron-like learning rule.

Additional resources

A key issue not considered here is how to handle **semantic underspecification**: cases in which there are multiple semantic interpretations for a single syntactic structure. Quantifier scope ambiguity is a classic example. Blackburn and Bos (2005) enumerate a number of approaches to this issue, and also provide links between natural language semantics and computational inference techniques. Much of the contemporary research on semantic parsing uses the framework of combinatory categorial grammar (CCG). Carpenter (1997) provides a comprehensive treatment of how CCG can support compositional semantic analysis. Another recent area of research is the semantics of multi-sentence texts. This can be handled with models of **dynamic semantics**, such as dynamic predicate logic (Groenendijk and Stokhof, 1991).

Alternative readings on formal semantics include an “informal” reading from Levy and Manning (2009), and a more involved introduction from Briscoe (2011). To learn more about ongoing research on data-driven semantic parsing, readers may consult the survey

6910 article by Liang and Potts (2015), tutorial slides and videos by Artzi and Zettlemoyer
 6911 (2013),¹² and the source code by Yoav Artzi¹³ and Percy Liang.¹⁴

6912 Exercises

- 6913 1. The **modus ponens** inference rule states that if we know $\phi \Rightarrow \psi$ and ϕ , then ψ must
 6914 be true. Justify this rule, using the definition of the \Rightarrow operator and some of the laws
 6915 provided in § 12.2.1, plus one additional identity: $\perp \vee \phi = \phi$.
- 6916 2. Convert the following examples into first-order logic, using the relations CAN-SLEEP,
 6917 MAKES-NOISE, and BROTHER.
 - 6918 • If Abigail makes noise, no one can sleep.
 - 6919 • If Abigail makes noise, someone cannot sleep.
 - 6920 • None of Abigail’s brothers can sleep.
 - 6921 • If one of Abigail’s brothers makes noise, Abigail cannot sleep.
- 6922 3. Extend the grammar fragment G_1 to include the ditransitive verb *teaches* and the
 6923 proper noun *Swahili*. Show how to derive the interpretation for the sentence *Alex*
 6924 *teaches Brit Swahili*, which should be $\text{TEACHES}(\text{ALEX}, \text{BRIT}, \text{SWAHILI})$. The grammar
 6925 need not be in Chomsky Normal Form. For the ditransitive verb, use NP_1 and NP_2
 6926 to indicate the two direct objects.
- 6927 4. Derive the semantic interpretation for the sentence *Alex likes every dog*, using gram-
 6928 mar fragment G_2 .
- 6929 5. Extend the grammar fragment G_2 to handle adjectives, so that the meaning of *an
 6930 angry dog* is $\lambda P. \exists x \text{DOG}(x) \wedge \text{ANGRY}(x) \wedge P(x)$. Specifically, you should supply the
 6931 lexical entry for the adjective *angry*, and you should specify the syntactic-semantic
 6932 productions $\text{NP} \rightarrow \text{DET } \text{NOM}$, $\text{NOM} \rightarrow \text{JJ } \text{NOM}$, and $\text{NOM} \rightarrow \text{NN}$.
- 6933 6. Extend your answer to the previous question to cover copula constructions with
 6934 predicative adjectives, such as *Alex is angry*. The interpretation should be $\text{ANGRY}(\text{ALEX})$.
 6935 You should add a verb phrase production $\text{VP} \rightarrow \text{V}_{\text{cop}} \text{JJ}$, and a terminal production
 6936 $\text{V}_{\text{cop}} \rightarrow \text{is}$. Show why your grammar extensions result in the correct interpretation.
- 6937 7. In Figure 12.5 and Figure 12.6, we treat the plurals *shoots* and *leaves* as entities. Revise
 6938 G_2 so that the interpretation of *Alex eats leaves* is $\forall x. (\text{LEAF}(x) \Rightarrow \text{EATS}(\text{ALEX}, x))$, and
 6939 show the resulting perceptron update.

¹²Videos are currently available at <http://yoavartzi.com/tutorial/>

¹³<http://yoavartzi.com/spf>

¹⁴<https://github.com/percyliang/sempre>

8. Statements like *every student eats a pizza* have two possible interpretations, depending on quantifier scope:

$$\forall x \exists y \text{PIZZA}(y) \wedge (\text{STUDENT}(x) \Rightarrow \text{EATS}(x, y)) \quad [12.32]$$

$$\exists y \forall x \text{PIZZA}(y) \wedge (\text{STUDENT}(x) \Rightarrow \text{EATS}(x, y)) \quad [12.33]$$

- 6940 a) Explain why these interpretations really are different.
- 6941 b) Which is generated by grammar G_2 ? Note that you may have to manipulate
6942 the logical form to exactly align with the grammar.
- 6943 9. *Modify G_2 so that produces the second interpretation in the previous problem.
6944 **Hint:** one possible solution involves changing the semantics of the sentence pro-
6945 duction and one other production.
- 6946 10. In the GeoQuery domain, give a natural language query that has multiple plausible
6947 semantic interpretations with the same denotation. List both interpretations and the
6948 denotation.
- 6949 **Hint:** There are many ways to do this, but one approach involves using toponyms
6950 (place names) that could plausibly map to several different entities in the model.

6951

Chapter 13

6952

Predicate-argument semantics

6953 This chapter considers more “lightweight” semantic representations, which discard some
6954 aspects of first-order logic, but focus on predicate-argument structures. Let’s begin by
6955 thinking about the semantics of events, with a simple example:

6956 (13.1) Asha gives Boyang a book.

6957 A first-order logical representation of this sentence is,

$$\exists x. \text{BOOK}(x) \wedge \text{GIVE}(\text{ASHA}, \text{BOYANG}, x) \quad [13.1]$$

6958 In this representation, we define variable x for the book, and we link the strings *Asha* and
6959 *Boyang* to entities ASHA and BOYANG. Because the action of giving involves a giver, a
6960 recipient, and a gift, the predicate GIVE must take three arguments.

6961 Now suppose we have additional information about the event:

6962 (13.2) Yesterday, Asha reluctantly gave Boyang a book.

6963 One possible solution is to extend the predicate GIVE to take additional arguments,

$$\exists x. \text{BOOK}(x) \wedge \text{GIVE}(\text{ASHA}, \text{BOYANG}, x, \text{YESTERDAY}, \text{RELUCTANTLY}) \quad [13.2]$$

But this is clearly unsatisfactory: *yesterday* and *reluctantly* are optional arguments, and we would need a different version of the GIVE predicate for every possible combination of arguments. **Event semantics** solves this problem by **reifying** the event as an existentially quantified variable e ,

$$\begin{aligned} \exists e, x. & \text{GIVE-EVENT}(e) \wedge \text{GIVER}(e, \text{ASHA}) \wedge \text{GIFT}(e, x) \wedge \text{BOOK}(e, x) \wedge \text{RECIPIENT}(e, \text{BOYANG}) \\ & \wedge \text{TIME}(e, \text{YESTERDAY}) \wedge \text{MANNER}(e, \text{RELUCTANTLY}) \end{aligned}$$

6964 In this way, each argument of the event — the giver, the recipient, the gift — can be rep-
 6965 resented with a relation of its own, linking the argument to the event e . The expression
 6966 GIVER(e , ASHA) says that ASHA plays the **role** of GIVER in the event. This reformulation
 6967 handles the problem of optional information such as the time or manner of the event,
 6968 which are called **adjuncts**. Unlike arguments, adjuncts are not a mandatory part of the
 6969 relation, but under this representation, they can be expressed with additional logical rela-
 6970 tions that are conjoined to the semantic interpretation of the sentence.¹

6971 The event semantic representation can be applied to nested clauses, e.g.,

6972 (13.3) Chris sees Asha pay Boyang.

This is done by using the event variable as an argument:

$$\exists e_1 \exists e_2 \text{SEE-EVENT}(e_1) \wedge \text{SEER}(e_1, \text{CHRIS}) \wedge \text{SIGHT}(e_1, e_2) \\ \wedge \text{PAY-EVENT}(e_2) \wedge \text{PAYER}(e_2, \text{ASHA}) \wedge \text{PAYEE}(e_2, \text{BOYANG}) \quad [13.3]$$

6973 As with first-order logic, the goal of event semantics is to provide a representation that
 6974 generalizes over many surface forms. Consider the following paraphrases of (13.1):

- 6975 (13.4) a. Asha gives a book to Boyang.
 6976 b. A book is given to Boyang by Asha.
 6977 c. A book is given by Asha to Boyang.
 6978 d. The gift of a book from Asha to Boyang ...

6979 All have the same event semantic meaning as Equation 13.1, but the ways in which the
 6980 meaning can be expressed are diverse. The final example does not even include a verb:
 6981 events are often introduced by verbs, but as shown by (13.4d), the noun *gift* can introduce
 6982 the same predicate, with the same accompanying arguments.

6983 **Semantic role labeling** (SRL) is a relaxed form of semantic parsing, in which each
 6984 semantic role is filled by a set of tokens from the text itself. This is sometimes called
 6985 “shallow semantics” because, unlike model-theoretic semantic parsing, role fillers need
 6986 not be symbolic expressions with denotations in some world model. A semantic role
 6987 labeling system is required to identify all predicates, and then specify the spans of text
 6988 that fill each role. To give a sense of the task, here is a more complicated example:

- 6989 (13.5) Boyang wants Asha to give him a linguistics book.

¹This representation is often called **Neo-Davidsonian event semantics**. The use of existentially-quantified event variables was proposed by Davidson (1967) to handle the issue of optional adjuncts. In Neo-Davidsonian semantics, this treatment of adjuncts is extended to mandatory arguments as well (e.g., Parsons, 1990).

6990 In this example, there are two predicates, expressed by the verbs *want* and *give*. Thus, a
 6991 semantic role labeler might return the following output:

- 6992 • (PREDICATE : *wants*, WANTED : *Boyang*, DESIRE : *Asha to give him a linguistics book*)
 6993 • (PREDICATE : *give*, GIVER : *Asha*, RECIPIENT : *him*, GIFT : *a linguistics book*)

6994 *Boyang* and *him* may refer to the same person, but the semantic role labeling is not re-
 6995 quired to resolve this reference. Other predicate-argument representations, such as **Ab-**
 6996 **stract Meaning Representation (AMR)**, do require reference resolution. We will return to
 6997 AMR in § 13.3, but first, let us further consider the definition of semantic roles.

6998 13.1 Semantic roles

6999 In event semantics, it is necessary to specify a number of additional logical relations to
 7000 link arguments to events: GIVER, RECIPIENT, SEER, SIGHT, etc. Indeed, every predicate re-
 7001 quires a set of logical relations to express its own arguments. In contrast, adjuncts such as
 7002 TIME and MANNER are shared across many types of events. A natural question is whether
 7003 it is possible to treat mandatory arguments more like adjuncts, by identifying a set of
 7004 generic argument types that are shared across many event predicates. This can be further
 7005 motivated by examples involving related verbs:

- 7006 (13.6) a. Asha gave Boyang a book.
 7007 b. Asha loaned Boyang a book.
 7008 c. Asha taught Boyang a lesson.
 7009 d. Asha gave Boyang a lesson.

7010 The respective roles of Asha, Boyang, and the book are nearly identical across the first
 7011 two examples. The third example is slightly different, but the fourth example shows that
 7012 the roles of GIVER and TEACHER can be viewed as related.

7013 One way to think about the relationship between roles such as GIVER and TEACHER is
 7014 by enumerating the set of properties that an entity typically possesses when it fulfills these
 7015 roles: givers and teachers are usually **animate** (they are alive and sentient) and **volitional**
 7016 (they choose to enter into the action).² In contrast, the thing that gets loaned or taught is
 7017 usually not animate or volitional; furthermore, it is unchanged by the event.

7018 Building on these ideas, **thematic roles** generalize across predicates by leveraging the
 7019 shared semantic properties of typical role fillers (Fillmore, 1968). For example, in exam-
 7020 ples (13.6a-13.6d), Asha plays a similar role in all four sentences, which we will call the

²There are always exceptions. For example, in the sentence *The C programming language has taught me a lot about perseverance*, the “teacher” is the *The C programming language*, which is presumably not animate or volitional.

	<i>Asha</i>	<i>gave</i>	<i>Boyang</i>	<i>a book</i>
VerbNet	AGENT		RECIPIENT	THEME
PropBank	ARG0: giver		ARG2: entity given to	ARG1: thing given
FrameNet	DONOR		RECIPIENT	THEME
	<i>Asha</i>	<i>taught</i>	<i>Boyang</i>	<i>algebra</i>
VerbNet	AGENT		RECIPIENT	TOPIC
PropBank	ARG0: teacher		ARG2: student	ARG1: subject
FrameNet	TEACHER		STUDENT	SUBJECT

Figure 13.1: Example semantic annotations according to VerbNet, PropBank, and FrameNet

7021 **agent.** This reflects several shared semantic properties: she is the one who is actively and
 7022 intentionally performing the action, while Boyang is a more passive participant; the book
 7023 and the lesson would play a different role, as non-animate participants in the event.

7024 Example annotations from three well known systems are shown in Figure 13.1. We
 7025 will now discuss these systems in more detail.

7026 13.1.1 VerbNet

7027 **VerbNet** (Kipper-Schuler, 2005) is a lexicon of verbs, and it includes thirty “core” thematic
 7028 roles played by arguments to these verbs. Here are some example roles, accompanied by
 7029 their definitions from the VerbNet Guidelines.³

- 7030 • AGENT: “ACTOR in an event who initiates and carries out the event intentionally or
 7031 consciously, and who exists independently of the event.”
- 7032 • PATIENT: “UNDERGOER in an event that experiences a change of state, location or
 7033 condition, that is causally involved or directly affected by other participants, and
 7034 exists independently of the event.”
- 7035 • RECIPIENT: “DESTINATION that is animate”
- 7036 • THEME: “UNDERGOER that is central to an event or state that does not have control
 7037 over the way the event occurs, is not structurally changed by the event, and/or is
 7038 characterized as being in a certain position or condition throughout the state.”
- 7039 • TOPIC: “THEME characterized by information content transferred to another partic-
 7040 ipant.”

³http://verbs.colorado.edu/verb-index/VerbNet_Guidelines.pdf

7041 VerbNet roles are organized in a hierarchy, so that a TOPIC is a type of THEME, which in
7042 turn is a type of UNDERGOER, which is a type of PARTICIPANT, the top-level category.

7043 In addition, VerbNet organizes verb senses into a class hierarchy, in which verb senses
7044 that have similar meanings are grouped together. Recall from § 4.2 that multiple meanings
7045 of the same word are called **senses**, and that WordNet identifies senses for many English
7046 words. VerbNet builds on WordNet, so that verb classes are identified by the WordNet
7047 senses of the verbs that they contain. For example, the verb class give-13.1 includes
7048 the first WordNet sense of *loan* and the second WordNet sense of *lend*.

7049 Each VerbNet class or subclass takes a set of thematic roles. For example, give-13.1
7050 takes arguments with the thematic roles of AGENT, THEME, and RECIPIENT;⁴ the pred-
7051 icate TEACH takes arguments with the thematic roles AGENT, TOPIC, RECIPIENT, and
7052 SOURCE.⁵ So according to VerbNet, *Asha* and *Boyang* play the roles of AGENT and RECIP-
7053 IENT in the sentences,

- 7054 (13.7) a. Asha gave Boyang a book.
7055 b. Asha taught Boyang algebra.

7056 The *book* and *algebra* are both THEMES, but *algebra* is a subcategory of THEME — a TOPIC
7057 — because it consists of information content that is given to the receiver.

7058 13.1.2 Proto-roles and PropBank

7059 Detailed thematic role inventories of the sort used in VerbNet are not universally accepted.
7060 For example, Dowty (1991, pp. 547) notes that “Linguists have often found it hard to agree
7061 on, and to motivate, the location of the boundary between role types.” He argues that a
7062 solid distinction can be identified between just two **proto-roles**:

7063 **Proto-Agent.** Characterized by volitional involvement in the event or state; sentience
7064 and/or perception; causing an event or change of state in another participant; move-
7065 ment; exists independently of the event.

7066 **Proto-Patient.** Undergoes change of state; causally affected by another participant; sta-
7067 tionary relative to the movement of another participant; does not exist indepen-
7068 dently of the event.⁶

⁴<https://verbs.colorado.edu/verb-index/vn/give-13.1.php>

⁵https://verbs.colorado.edu/verb-index/vn/transfer_mesg-37.1.1.php

⁶Reisinger et al. (2015) ask crowd workers to annotate these properties directly, finding that annotators tend to agree on the properties of each argument. They also find that in English, arguments having more proto-agent properties tend to appear in subject position, while arguments with more proto-patient properties appear in object position.

7069 In the examples in Figure 13.1, Asha has most of the proto-agent properties: in giving
 7070 the book to Boyang, she is acting volitionally (as opposed to *Boyang got a book from Asha*, in
 7071 which it is not clear whether Asha gave up the book willingly); she is sentient; she causes a
 7072 change of state in Boyang; she exists independently of the event. Boyang has some proto-
 7073 agent properties: he is sentient and exists independently of the event. But he also has
 7074 some proto-patient properties: he is the one who is causally affected and who undergoes
 7075 change of state. The book that Asha gives Boyang has even fewer of the proto-agent
 7076 properties: it is not volitional or sentient, and it has no causal role. But it also lacks many
 7077 of the proto-patient properties: it does not undergo change of state, exists independently
 7078 of the event, and is not stationary.

7079 The **Proposition Bank**, or PropBank (Palmer et al., 2005), builds on this basic agent-
 7080 patient distinction, as a middle ground between generic thematic roles and roles that are
 7081 specific to each predicate. Each verb is linked to a list of numbered arguments, with ARG0
 7082 as the proto-agent and ARG1 as the proto-patient. Additional numbered arguments are
 7083 verb-specific. For example, for the predicate TEACH,⁷ the arguments are:

- 7084 • ARG0: the teacher
- 7085 • ARG1: the subject
- 7086 • ARG2: the student(s)

7087 Verbs may have any number of arguments: for example, WANT and GET have five, while
 7088 EAT has only ARG0 and ARG1. In addition to the semantic arguments found in the frame
 7089 files, roughly a dozen general-purpose adjuncts may be used in combination with any
 7090 verb. These are shown in Table 13.1.

7091 PropBank-style semantic role labeling is annotated over the entire Penn Treebank. This
 7092 annotation includes the sense of each verbal predicate, as well as the argument spans.

7093 13.1.3 FrameNet

7094 Semantic **frames** are descriptions of situations or events. Frames may be *evoked* by one
 7095 of their **lexical units** (often a verb, but not always), and they include some number of
 7096 **frame elements**, which are like roles (Fillmore, 1976). For example, the act of teaching
 7097 is a frame, and can be evoked by the verb *taught*; the associated frame elements include
 7098 the teacher, the student(s), and the subject being taught. Frame semantics has played a
 7099 significant role in the history of artificial intelligence, in the work of Minsky (1974) and
 7100 Schank and Abelson (1977). In natural language processing, the theory of frame semantics
 7101 has been implemented in **FrameNet** (Fillmore and Baker, 2009), which consists of a lexicon

⁷<http://verbs.colorado.edu/propbank/framesets-english-aliases/teach.html>

TMP	time	<i>Boyang ate a bagel</i> [AM-TMP <i>yesterday</i>].
LOC	location	<i>Asha studies in</i> [AM-LOC <i>Stuttgart</i>]
MOD	modal verb	<i>Asha</i> [AM-MOD <i>will</i>] <i>study in Stuttgart</i>
ADV	general purpose	[AM-ADV <i>Luckily</i>], <i>Asha knew algebra</i> .
MNR	manner	<i>Asha ate</i> [AM-MNR <i>aggressively</i>].
DIS	discourse connective	[AM-DIS <i>However</i>], <i>Asha prefers algebra</i> .
PRP	purpose	<i>Barry studied</i> [AM-PRP <i>to pass the bar</i>].
DIR	direction	<i>Workers dumped burlap sacks</i> [AM-DIR <i>into a bin</i>].
NEG	negation	<i>Asha does</i> [AM-NEG <i>not</i>] <i>speak Albanian</i> .
EXT	extent	<i>Prices increased</i> [AM-EXT <i>4%</i>].
CAU	cause	<i>Boyang returned the book</i> [AM-CAU <i>because it was overdue</i>].

Table 13.1: PropBank adjuncts (Palmer et al., 2005), sorted by frequency in the corpus

7102 of roughly 1000 frames, and a corpus of more than 200,000 “exemplar sentences,” in which
 7103 the frames and their elements are annotated.⁸

7104 Rather than seeking to link semantic roles such as TEACHER and GIVER into the-
 7105 matic roles such as AGENT, FrameNet aggressively groups verbs into frames, and links
 7106 semantically-related roles across frames. For example, the following two sentences would
 7107 be annotated identically in FrameNet:

- 7108 (13.8) a. Asha taught Boyang algebra.
 7109 b. Boyang learned algebra from Asha.

7110 This is because *teach* and *learn* are both lexical units in the EDUCATION-TEACHING frame.
 7111 Furthermore, roles can be shared even when the frames are distinct, as in the following
 7112 two examples:

- 7113 (13.9) a. Asha gave Boyang a book.
 7114 b. Boyang got a book from Asha.

7115 The GIVING and GETTING frames both have RECIPIENT and THEME elements, so Boyang
 7116 and the book would play the same role. Asha’s role is different: she is the DONOR in the
 7117 GIVING frame, and the SOURCE in the GETTING frame. FrameNet makes extensive use of
 7118 multiple inheritance to share information across frames and frame elements: for example,
 7119 the COMMERCE-SELL and LENDING frames inherit from GIVING frame.

⁸Current details and data can be found at <https://framenet.icsi.berkeley.edu/>

7120 13.2 Semantic role labeling

7121 The task of semantic role labeling is to identify the parts of the sentence comprising the
 7122 semantic roles. In English, this task is typically performed on the PropBank corpus, with
 7123 the goal of producing outputs in the following form:

7124 (13.10) [ARG0 Asha] [GIVE.01 gave] [ARG2 Boyang's mom] [ARG1 a book] [AM-TMP yesterday].

7125 Note that a single sentence may have multiple verbs, and therefore a given word may be
 7126 part of multiple role-fillers:

7127 (13.11) [ARG0 Asha] [WANT.01 wanted]
 Asha wanted

7128 [ARG1 Boyang to give her the book].
 [ARG0 Boyang] [GIVE.01 to give] [ARG2 her] [ARG1 the book].

7129 13.2.1 Semantic role labeling as classification

7130 PropBank is annotated on the Penn Treebank, and annotators used phrasal constituents
 7131 (\S 9.2.2) to fill the roles. PropBank semantic role labeling can be viewed as the task of as-
 7132 signing to each phrase a label from the set $\mathcal{R} = \{\emptyset, \text{PRED}, \text{ARG0}, \text{ARG1}, \text{ARG2}, \dots, \text{AM-LOC}, \text{AM-TMP}, \dots\}$
 7133 with respect to each predicate. If we treat semantic role labeling as a classification prob-
 7134 lem, we obtain the following functional form:

$$\hat{y}_{(i,j)} = \underset{y}{\operatorname{argmax}} \psi(\mathbf{w}, y, i, j, \rho, \tau), \quad [13.4]$$

7135 where,

- 7136 • (i, j) indicates the span of a phrasal constituent $(w_{i+1}, w_{i+2}, \dots, w_j)$;⁹
- 7137 • \mathbf{w} represents the sentence as a sequence of tokens;
- 7138 • ρ is the index of the predicate verb in \mathbf{w} ;
- 7139 • τ is the structure of the phrasal constituent parse of \mathbf{w} .

7140 Early work on semantic role labeling focused on discriminative feature-based models,
 7141 where $\psi(\mathbf{w}, y, i, j, \rho, \tau) = \theta \cdot f(\mathbf{w}, y, i, j, \rho, \tau)$. Table 13.2 shows the features used in a sem-
 7142 inal paper on FrameNet semantic role labeling (Gildea and Jurafsky, 2002). By 2005 there

⁹PropBank roles can also be filled by **split constituents**, which are discontinuous spans of text. This situation most frequently in reported speech, e.g. [ARG1 *By addressing these problems*], *Mr. Maxwell said*, [ARG1 *the new funds have become extremely attractive.*] (example adapted from Palmer et al., 2005). This issue is typically addressed by defining “continuation arguments”, e.g. C-ARG1, which refers to the continuation of ARG1 after the split.

Predicate lemma and POS tag	The lemma of the predicate verb and its part-of-speech tag
Voice	Whether the predicate is in active or passive voice, as determined by a set of syntactic patterns for identifying passive voice constructions
Phrase type	The constituent phrase type for the proposed argument in the parse tree, e.g. NP, PP
Headword and POS tag	The head word of the proposed argument and its POS tag, identified using the Collins (1997) rules
Position	Whether the proposed argument comes before or after the predicate in the sentence
Syntactic path	The set of steps on the parse tree from the proposed argument to the predicate (described in detail in the text)
Subcategorization	The syntactic production from the first branching node above the predicate. For example, in Figure 13.2, the subcategorization feature around <i>taught</i> would be VP → VBD NP PP.

Table 13.2: Features used in semantic role labeling by Gildea and Jurafsky (2002).

7143 were several systems for PropBank semantic role labeling, and their approaches and fea-
 7144 ture sets are summarized by Carreras and Màrquez (2005). Typical features include: the
 7145 phrase type, head word, part-of-speech, boundaries, and neighbors of the proposed argu-
 7146 ment $w_{i+1:j}$; the word, lemma, part-of-speech, and voice of the verb w_ρ (active or passive),
 7147 as well as features relating to its frameset; the distance and path between the verb and
 7148 the proposed argument. In this way, semantic role labeling systems are high-level “con-
 7149 sumers” in the NLP stack, using features produced from lower-level components such as
 7150 part-of-speech taggers and parsers. More comprehensive feature sets are enumerated by
 7151 Das et al. (2014) and Täckström et al. (2015).

7152 A particularly powerful class of features relate to the **syntactic path** between the ar-
 7153 gument and the predicate. These features capture the sequence of moves required to get
 7154 from the argument to the verb by traversing the phrasal constituent parse of the sentence.
 7155 The idea of these features is to capture syntactic regularities in how various arguments
 7156 are realized. Syntactic path features are best illustrated by example, using the parse tree
 7157 in Figure 13.2:

- 7158 • The path from *Asha* to the verb *taught* is NNP↑NP↑S↓VP↓VBD. The first part of
 7159 the path, NNP↑NP↑S, means that we must travel up the parse tree from the NNP
 7160 tag (proper noun) to the S (sentence) constituent. The second part of the path,
 7161 S↓VP↓VBD, means that we reach the verb by producing a VP (verb phrase) from

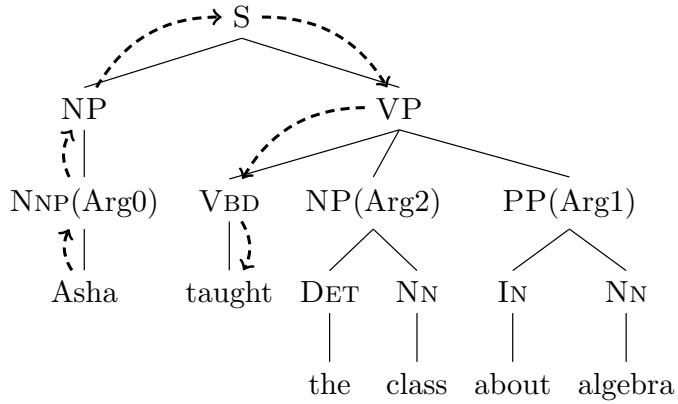


Figure 13.2: Semantic role labeling on the phrase-structure parse tree for a sentence. The dashed line indicates the syntactic path from *Asha* to the predicate verb *taught*.

7162 the S constituent, and then by producing a VBD (past tense verb). This feature is
 7163 consistent with *Asha* being in subject position, since the path includes the sentence
 7164 root S.

- 7165 • The path from *the class* to *taught* is NP↑VP↓VBD. This is consistent with *the class*
 7166 being in object position, since the path passes through the VP node that dominates
 7167 the verb *taught*.

7168 Because there are many possible path features, it can also be helpful to look at smaller
 7169 parts: for example, the upward and downward parts can be treated as separate features;
 7170 another feature might consider whether S appears anywhere in the path.

7171 Rather than using the constituent parse, it is also possible to build features from the **de-**
 7172 **pendency path** (see § 11.4) between the head word of each argument and the verb (Prad-
 7173 han et al., 2005). Using the Universal Dependency part-of-speech tagset and dependency
 7174 relations (Nivre et al., 2016), the dependency path from *Asha* to *taught* is PROPN $\xleftarrow[\text{NSUBJ}]{} \text{VERB}$,
 7175 because *taught* is the head of a relation of type $\xleftarrow[\text{NSUBJ}]{} \text{VERB}$. Similarly, the dependency
 7176 path from *class* to *taught* is NOUN $\xleftarrow[\text{DOBJ}]{} \text{VERB}$, because *class* heads the noun phrase that is a
 7177 direct object of *taught*. A more interesting example is *Asha wanted to teach the class*, where
 7178 the path from *Asha* to *teach* is PROPN $\xleftarrow[\text{NSUBJ}]{} \text{VERB} \rightarrow \text{VERB}$. The right-facing arrow in sec-
 7179 ond relation indicates that *wanted* is the head of its XCOMP relation with *teach*.

7180 **13.2.2 Semantic role labeling as constrained optimization**

7181 A potential problem with treating SRL as a classification problem is that there are a num-
 7182 ber of sentence-level **constraints**, which a classifier might violate.

- 7183 • For a given verb, there can be only one argument of each type (ARG0, ARG1, etc.)
 7184 • Arguments cannot overlap. This problem arises when we are labeling the phrases
 7185 in a constituent parse tree, as shown in Figure 13.2: if we label the PP *about algebra*
 7186 as an argument or adjunct, then its children *about* and *algebra* must be labeled as \emptyset .
 7187 The same constraint also applies to the syntactic ancestors of this phrase.

7188 These constraints introduce dependencies across labeling decisions. In structure pre-
 7189 diction problems such as sequence labeling and parsing, such dependencies are usually
 7190 handled by defining a scoring over the entire structure, \mathbf{y} . Efficient inference requires
 7191 that the global score decomposes into local parts: for example, in sequence labeling, the
 7192 scoring function decomposes into scores of pairs of adjacent tags, permitting the applica-
 7193 tion of the Viterbi algorithm for inference. But the constraints that arise in semantic role
 7194 labeling are less amenable to local decomposition.¹⁰ We therefore consider **constrained**
 7195 **optimization** as an alternative solution.

Let the set $\mathcal{C}(\tau)$ refer to all labelings that obey the constraints introduced by the parse τ . The semantic role labeling problem can be reformulated as a constrained optimization over $\mathbf{y} \in \mathcal{C}(\tau)$,

$$\begin{aligned} \max_{\mathbf{y}} \quad & \sum_{(i,j) \in \tau} \psi(\mathbf{w}, y_{i,j}, i, j, \rho, \tau) \\ \text{s.t.} \quad & \mathbf{y} \in \mathcal{C}(\tau). \end{aligned} \quad [13.5]$$

7196 In this formulation, the objective (shown on the first line) is a separable function of each
 7197 individual labeling decision, but the constraints (shown on the second line) apply to the
 7198 overall labeling. The sum $\sum_{(i,j) \in \tau}$ indicates that we are summing over all constituent
 7199 spans in the parse τ . The expression s.t. in the second line means that we maximize the
 7200 objective *subject to* the constraint $\mathbf{y} \in \mathcal{C}(\tau)$.

7201 A number of practical algorithms exist for restricted forms of constrained optimiza-
 7202 tion. One such restricted form is **integer linear programming**, in which the objective and
 7203 constraints are linear functions of integer variables. To formulate SRL as an integer linear
 7204 program, we begin by rewriting the labels as a set of binary variables $\mathbf{z} = \{z_{i,j,r}\}$ (Pun-
 7205 yakanok et al., 2008),

$$z_{i,j,r} = \begin{cases} 1, & y_{i,j} = r \\ 0, & \text{otherwise,} \end{cases} \quad [13.6]$$

¹⁰Dynamic programming solutions have been proposed by Tromble and Eisner (2006) and Täckström et al. (2015), but they involve creating a trellis structure whose size is exponential in the number of labels.

7206 where $r \in \mathcal{R}$ is a label in the set $\{\text{ARG0}, \text{ARG1}, \dots, \text{AM-LOC}, \dots, \emptyset\}$. Thus, the variables
 7207 z are a binarized version of the semantic role labeling y .

The objective can then be formulated as a linear function of z .

$$\sum_{(i,j) \in \tau} \psi(\mathbf{w}, y_{i,j}, i, j, \rho, \tau) = \sum_{i,j,r} \psi(\mathbf{w}, r, i, j, \rho, \tau) \times z_{i,j,r}, \quad [13.7]$$

7208 which is the sum of the scores of all relations, as indicated by $z_{i,j,r}$.

Constraints Integer linear programming permits linear inequality constraints, of the general form $\mathbf{A}z \leq \mathbf{b}$, where the parameters \mathbf{A} and \mathbf{b} define the constraints. To make this more concrete, let's start with the constraint that each non-null role type can occur only once in a sentence. This constraint can be written,

$$\forall r \neq \emptyset, \quad \sum_{(i,j) \in \tau} z_{i,j,r} \leq 1. \quad [13.8]$$

7209 Recall that $z_{i,j,r} = 1$ iff the span (i, j) has label r ; this constraint says that for each possible
 7210 label $r \neq \emptyset$, there can be at most one (i, j) such that $z_{i,j,r} = 1$. Rewriting this constraint
 7211 can be written in the form $\mathbf{A}z \leq \mathbf{b}$, as you will find if you complete the exercises at the
 7212 end of the chapter.

Now consider the constraint that labels cannot overlap. Let's define the convenience function $o((i, j), (i', j')) = 1$ iff (i, j) overlaps (i', j') , and zero otherwise. Thus, o will indicate if a constituent (i', j') is either an ancestor or descendant of (i, j) . The constraint is that if two constituents overlap, only one can have a non-null label:

$$\forall (i, j) \in \tau, \quad \sum_{(i', j') \in \tau} \sum_{r \neq \emptyset} o((i, j), (i', j')) \times z_{i',j',r} \leq 1, \quad [13.9]$$

7213 where $o((i, j), (i, j)) = 1$.

In summary, the semantic role labeling problem can thus be rewritten as the following integer linear program,

$$\max_{z \in \{0,1\}^{|\tau|}} \quad \sum_{(i,j) \in \tau} \sum_{r \in \mathcal{R}} z_{i,j,r} \psi_{i,j,r} \quad [13.10]$$

$$s.t. \quad \forall r \neq \emptyset, \quad \sum_{(i,j) \in \tau} z_{i,j,r} \leq 1. \quad [13.11]$$

$$\forall (i, j) \in \tau, \quad \sum_{(i', j') \in \tau} \sum_{r \neq \emptyset} o((i, j), (i', j')) \times z_{i',j',r} \leq 1. \quad [13.12]$$

7214 **Learning with constraints** Learning can be performed in the context of constrained op-
 7215 timization using the usual perceptron or large-margin classification updates. Because
 7216 constrained inference is generally more time-consuming, a key question is whether it is
 7217 necessary to apply the constraints during learning. Chang et al. (2008) find that better per-
 7218 formance can be obtained by learning *without* constraints, and then applying constraints
 7219 only when using the trained model to predict semantic roles for unseen data.

7220 **How important are the constraints?** Das et al. (2014) find that an unconstrained, classification-
 7221 based method performs nearly as well as constrained optimization for FrameNet parsing:
 7222 while it commits many violations of the “no-overlap” constraint, the overall F_1 score is
 7223 less than one point worse than the score at the constrained optimum. Similar results
 7224 were obtained for PropBank semantic role labeling by Punyakanok et al. (2008). He et al.
 7225 (2017) find that constrained inference makes a bigger impact if the constraints are based
 7226 on manually-labeled “gold” syntactic parses. This implies that errors from the syntac-
 7227 tic parser may limit the effectiveness of the constraints. Punyakanok et al. (2008) hedge
 7228 against parser error by including constituents from several different parsers; any con-
 7229 stituent can be selected from any parse, and additional constraints ensure that overlap-
 7230 ping constituents are not selected.

7231 **Implementation** Integer linear programming solvers such as `glpk`,¹¹ `cplex`,¹² and `Gurobi`¹³
 7232 allow inequality constraints to be expressed directly in the problem definition, rather than
 7233 in the matrix form $\mathbf{A}z \leq \mathbf{b}$. The time complexity of integer linear programming is theoreti-
 7234 cally exponential in the number of variables $|z|$, but in practice these off-the-shelf solvers
 7235 obtain good solutions efficiently. Using a standard desktop computer, Das et al. (2014)
 7236 report that the `cplex` solver requires 43 seconds to perform inference on the FrameNet
 7237 test set, which contains 4,458 predicates.

7238 Recent work has shown that many constrained optimization problems in natural lan-
 7239 guage processing can be solved in a highly parallelized fashion, using optimization tech-
 7240 niques such as **dual decomposition**, which are capable of exploiting the underlying prob-
 7241 lem structure (Rush et al., 2010). Das et al. (2014) apply this technique to FrameNet se-
 7242 mantic role labeling, obtaining an order-of-magnitude speedup over `cplex`.

7243 13.2.3 Neural semantic role labeling

7244 Neural network approaches to SRL have tended to treat it as a sequence labeling task,
 7245 using a labeling scheme such as the **BIO notation**, which we previously saw in named
 7246 entity recognition (§ 8.3). In this notation, the first token in a span of type ARG1 is labeled

¹¹<https://www.gnu.org/software/glpk/>

¹²<https://www-01.ibm.com/software/commerce/optimization/cplex-optimizer/>

¹³<http://www.gurobi.com/>

7247 B-ARG1; all remaining tokens in the span are *inside*, and are therefore labeled I-ARG1.
 7248 Tokens outside any argument are labeled O. For example:

- 7249 (13.12) *Asha taught Boyang 's mom about algebra*
 B-ARG0 PRED B-ARG2 I-ARG2 I-ARG2 B-ARG1 I-ARG1

Recurrent neural networks (§ 7.6) are a natural approach to this tagging task. For example, Zhou and Xu (2015) apply a deep bidirectional multilayer LSTM (see § 7.6) to PropBank semantic role labeling. In this model, each bidirectional LSTM serves as input for another, higher-level bidirectional LSTM, allowing complex non-linear transformations of the original input embeddings, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M]$. The hidden state of the final LSTM is $\mathbf{Z}^{(K)} = [z_1^{(K)}, z_2^{(K)}, \dots, z_M^{(K)}]$. The “emission” score for each tag $Y_m = y$ is equal to the inner product $\theta_y \cdot z_m^{(K)}$, and there is also a transition score for each pair of adjacent tags. The complete model can be written,

$$\mathbf{Z}^{(1)} = \text{BiLSTM}(\mathbf{X}) \quad [13.13]$$

$$\mathbf{Z}^{(i)} = \text{BiLSTM}(\mathbf{Z}^{(i-1)}) \quad [13.14]$$

$$\hat{\mathbf{y}} = \underset{\mathbf{y}}{\text{argmax}} \sum_{m=1}^M \Theta^{(y)} \mathbf{z}_m^{(K)} + \psi_{y_{m-1}, y_m}. \quad [13.15]$$

7250 Note that the final step maximizes over the entire labeling \mathbf{y} , and includes a score for
 7251 each tag transition ψ_{y_{m-1}, y_m} . This combination of LSTM and pairwise potentials on tags
 7252 is an example of an **LSTM-CRF**. The maximization over \mathbf{y} is performed by the Viterbi
 7253 algorithm.

7254 This model strongly outperformed alternative approaches at the time, including con-
 7255 strained decoding and convolutional neural networks.¹⁴ More recent work has combined
 7256 recurrent neural network models with constrained decoding, using the A^* search algo-
 7257 rithm to search over labelings that are feasible with respect to the constraints (He et al.,
 7258 2017). This yields small improvements over the method of Zhou and Xu (2015). He et al.
 7259 (2017) obtain larger improvements by creating an **ensemble** of SRL systems, each trained
 7260 on an 80% subsample of the corpus. The average prediction across this ensemble is more
 7261 robust than any individual model.

7262 13.3 Abstract Meaning Representation

7263 Semantic role labeling transforms the task of semantic parsing to a labeling task. Consider
 7264 the sentence,

¹⁴The successful application of **convolutional neural networks** to semantic role labeling by Collobert and Weston (2008) was an influential early result in the current wave of neural networks in natural language processing.

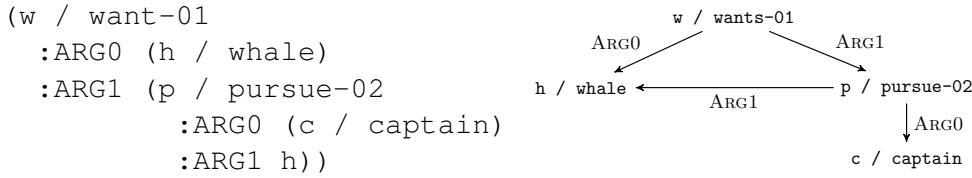


Figure 13.3: Two views of the AMR representation for the sentence *The whale wants the captain to pursue him.*

7265 (13.13) The whale wants the captain to pursue him.

7266 The PropBank semantic role labeling analysis is:

7267 • (PREDICATE : *wants*, ARG0 : *the whale*, ARG1 : *the captain to pursue him*)

7268 • (PREDICATE : *pursue*, ARG0 : *the captain*, ARG1 : *him*)

7269 The **Abstract Meaning Representation (AMR)** unifies this analysis into a graph structure, in which each node is a **variable**, and each edge indicates a **concept** (Banarescu et al., 2013). This can be written in two ways, as shown in Figure 13.3. On the left is the PENMAN notation (Matthiessen and Bateman, 1991), in which each set of parentheses introduces a variable. Each variable is an **instance** of a concept, which is indicated with the slash notation: for example, *w / want-01* indicates that the variable *w* is an instance of the concept *want-01*, which in turn refers to the PropBank frame for the first sense of the verb *want*; *pursue-02* refers to the second sense of *pursue*. Relations are introduced with colons: for example, *:ARG0 (c / captain)* indicates a relation of type *ARG0* with the newly-introduced variable *c*. Variables can be reused, so that when the variable *h* appears again as an argument to *p*, it is understood to refer to the same whale in both cases. This arrangement is indicated compactly in the graph structure on the right, with edges indicating concepts.

7270 One way in which AMR differs from PropBank-style semantic role labeling is that it
 7271 reifies each entity as a variable: for example, the *whale* in (13.13) is reified in the variable
 7272 *h*, which is reused as *ARG0* in its relationship with *w / want-01*, and as *ARG1* in its
 7273 relationship with *p / pursue-02*. Reifying entities as variables also makes it possible
 7274 to represent the substructure of noun phrases more explicitly. For example, *Asha borrowed*
 7275 *the algebra book* would be represented as:

7276 (b / borrow-01
 7277 :ARG0 (p / person
 7278 :name (n / name
 7279 :op1 "Asha"))

7292 :ARG1 (b2 / book
 7293 :topic (a / algebra))

7294 This indicates that the variable *p* is a person, whose name is the variable *n*; that name
 7295 has one token, the string *Asha*. Similarly, the variable *b2* is a book, and the *topic* of *b2*
 7296 is a variable *a* whose type is *algebra*. The relations *name* and *topic* are examples of
 7297 “non-core roles”, which are similar to adjunct modifiers in PropBank. However, AMR’s
 7298 inventory is more extensive, including more than 70 non-core roles, such as negation,
 7299 time, manner, frequency, and location. Lists and sequences — such as the list of tokens in
 7300 a name — are described using the roles *op1*, *op2*, etc.

7301 Another feature of AMR is that a semantic predicate can be introduced by any syntac-
 7302 tic element, as in the following examples from Banarescu et al. (2013):

- 7303 (13.14) a. The boy destroyed the room.
 7304 b. the destruction of the room by the boy ...
 7305 c. the boy’s destruction of the room ...

7306 All these examples have the same semantics in AMR,

7307 (d / destroy-01
 7308 :ARG0 (b / boy)
 7309 :ARG1 (r / room))

7310 The noun *destruction* is linked to the verb *destroy*, which is captured by the PropBank
 7311 frame *destroy-01*. This can happen with adjectives as well: in the phrase *the attractive*
 7312 *spy*, the adjective *attractive* is linked to the PropBank frame *attract-01*:

7313 (s / spy
 7314 :ARG0-of (a / attract-01))

7315 In this example, *ARG0-of* is an **inverse relation**, indicating that *s* is the *ARG0* of the
 7316 predicate *a*. Inverse relations make it possible for all AMR parses to have a single root
 7317 concept.

7318 While AMR goes farther than semantic role labeling, it does not link semantically-
 7319 related frames such as *buy*/*sell* (as FrameNet does). AMR also does not handle quanti-
 7320 fication (as first-order predicate calculus does), and it makes no attempt to handle noun
 7321 number and verb tense (as PropBank does).

13.3.1 AMR Parsing

Abstract Meaning Representation is not a labeling of the original text — unlike PropBank semantic role labeling, and most of the other tagging and parsing tasks that we have encountered thus far. The AMR for a given sentence may include multiple concepts for single words in the sentence: as we have seen, the sentence *Asha likes algebra* contains both person and name concepts for the word *Asha*. Conversely, words in the sentence may not appear in the AMR: in *Boyang made a tour of campus*, the light verb *make* would not appear in the AMR, which would instead be rooted on the predicate *tour*. As a result, AMR is difficult to parse, and even evaluating AMR parsing involves considerable algorithmic complexity (Cai and Yates, 2013).

A further complexity is that AMR labeled datasets do not explicitly show the alignment between the AMR annotation and the words in the sentence. For example, the link between the word *wants* and the concept *want-01* is not annotated. To acquire training data for learning-based parsers, it is therefore necessary to first perform an alignment between the training sentences and their AMR parses. Flanigan et al. (2014) introduce a rule-based parser, which links text to concepts through a series of increasingly high-recall steps.

As with dependency parsing, AMR can be parsed by graph-based methods that explore the space of graph structures, or by incremental transition-based algorithms. One approach to graph-based AMR parsing is to first group adjacent tokens into local substructures, and then to search the space of graphs over these substructures (Flanigan et al., 2014). The identification of concept subgraphs can be formulated as a sequence labeling problem, and the subsequent graph search can be solved using integer linear programming (§ 13.2.2). Various transition-based parsing algorithms have been proposed. Wang et al. (2015) construct an AMR graph by incrementally modifying the syntactic dependency graph. At each step, the parser performs an action: for example, adding an AMR relation label to the current dependency edge, swapping the direction of a syntactic dependency edge, or cutting an edge and reattaching the orphaned subtree to a new parent.

Additional resources

Practical semantic role labeling was first made possible by the PropBank annotations on the Penn Treebank (Palmer et al., 2005). Abend and Rappoport (2017) survey several semantic representation schemes, including semantic role labeling and AMR. Other linguistic features of AMR are summarized in the original paper (Banarescu et al., 2013) and the tutorial slides by Schneider et al. (2015). Recent shared tasks have undertaken semantic dependency parsing, in which the goal is to identify semantic relationships between pairs of words (Oepen et al., 2014); see Ivanova et al. (2012) for an overview of connections between syntactic and semantic dependencies.

7359 **Exercises**

7360 1. Write out an event semantic representation for the following sentences. You may
7361 make up your own predicates.

7362 (13.15) *Abigail shares with Max.*

7363 (13.16) *Abigail reluctantly shares a toy with Max.*

7364 (13.17) *Abigail hates to share with Max.*

7365 2. Find the PropBank framesets for *share* and *hate* at <http://verbs.colorado.edu/propbank/framesets-english-aliases/>, and rewrite your answers from the
7366 previous question, using the thematic roles ARG0, ARG1, and ARG2.
7367

7368 3. Compute the syntactic path features for Abigail and Max in each of the example sentences
7369 (13.15) and (13.17) in Question 1, with respect to the verb *share*. If you’re not
7370 sure about the parse, you can try an online parser such as <http://nlp.stanford.edu:8080/parser/>.
7371

7372 4. Compute the dependency path features for Abigail and Max in each of the example
7373 sentences (13.15) and (13.17) in Question 1, with respect to the verb *share*. Again, if
7374 you’re not sure about the parse, you can try an online parser such as <http://nlp.stanford.edu:8080/parser/>. As a hint, the dependency relation between *share*
7375 and *Max* is OBL according to the Universal Dependency treebank.
7376

7377 5. PropBank semantic role labeling includes **reference arguments**, such as,

7378 (13.18) [AM-LOC *The bed*] on [R-AM-LOC *which*] I slept broke.¹⁵

7379 The label R-AM-LOC indicates that the word *which* is a reference to *The bed*, which
7380 expresses the location of the event. Reference arguments must have referents: the
7381 tag R-AM-LOC can appear only when AM-LOC also appears in the sentence. Show
7382 how to express this as a linear constraint, specifically for the tag R-AM-LOC. Be sure
7383 to correctly handle the case in which neither AM-LOC nor R-AM-LOC appear in the
7384 sentence.

7385 6. Explain how to express the constraints on semantic role labeling in Equation 13.8
7386 and Equation 13.9 in the general form $Az \geq b$.

7387 7. Produce the AMR annotations for the following examples:

7388 (13.19) *The girl likes the boy.*

¹⁵Example from 2013 NAACL tutorial slides by Shumin Wu

- 7389 (13.20) The girl was liked by the boy.
 7390 (13.21) Abigail likes Maxwell Aristotle.
 7391 (13.22) The spy likes the attractive boy.
 7392 (13.23) The girl doesn't like the boy.
 7393 (13.24) The girl likes her dog.

7394 For (13.21), recall that multi-token names are created using `op1`, `op2`, etc. You will
 7395 need to consult Banarescu et al. (2013) for (13.23), and Schneider et al. (2015) for
 7396 (13.24). You may assume that *her* refers to *the girl* in this example.

- 7397 8. In this problem, you will build a FrameNet sense classifier for the verb *can*, which
 7398 can evoke two frames: POSSIBILITY (can you order a salad with french fries?) and
 7399 CAPABILITY (can you eat a salad with chopsticks?).

7400 To build the dataset, access the FrameNet corpus in NLTK:

```
7401 import nltk
7402 nltk.download('framenet_v17')
7403 from nltk.corpus import framenet as fn
```

7404 Next, find instances in which the lexical unit `can.v` (the verb form of *can*) evokes a
 7405 frame. Do this by iterating over `fn.docs()`, and then over sentences, and then

```
7406 for doc in fn.docs():
7407     if 'sentence' in doc:
7408         for sent in doc['sentence']:
7409             for anno_set in sent['annotationSet']:
7410                 if 'luName' in anno_set and anno_set['luName'] == 'can.v':
7411                     pass # your code here
```

7412 Use the field `frameName` as a label, and build a set of features from the field `text`.
 7413 Train a classifier to try to accurately predict the `frameName`, disregarding cases
 7414 other than CAPABILITY and POSSIBILITY. Treat the first hundred instances as a training
 7415 set, and the remaining instances as the test set. Can you do better than a classifier
 7416 that simply selects the most common class?

- 7417 9. *Download the PropBank sample data, using NLTK (<http://www.nltk.org/howto/propbank.html>).

- 7419 a) Use a deep learning toolkit such as PyTorch to train a BiLSTM sequence labeling
 7420 model (§ 7.6) to identify words or phrases that are predicates, e.g., *we/O*
 7421 *took/B-PRED a/I-PRED walk/I-PRED together/O*. Your model should compute
 7422 the tag score from the BiLSTM hidden state $\psi(y_m) = \beta_y \cdot h_m$.
- 7423 b) Optionally, implement Viterbi to improve the predictions of the model in the
 7424 previous section.

7425 c) Try to identify ARG0 and ARG1 for each predicate. You should again use the
 7426 BiLSTM and BIO notation, but you may want to include the BiLSTM hidden
 7427 state at the location of the predicate in your prediction model, e.g., $\psi(y_m) =$
 7428 $\beta_y \cdot [\mathbf{h}_m; \mathbf{h}_{\hat{r}}]$, where \hat{r} is the predicted location of the (first word of the) predicate.

7429 10. Using an off-the-shelf PropBank SRL system,¹⁶ build a simplified question answer-
 7430 ing system in the style of Shen and Lapata (2007). Specifically, your system should
 7431 do the following:

- 7432 • For each document in a collection, it should apply the semantic role labeler,
 7433 and should store the output as a tuple.
- 7434 • For a question, your system should again apply the semantic role labeler. If
 7435 any of the roles are filled by a *wh*-pronoun, you should mark that role as the
 7436 expected answer phrase (EAP).
- 7437 • To answer the question, search for a stored tuple which matches the question as
 7438 well as possible (same predicate, no incompatible semantic roles, and as many
 7439 matching roles as possible). Align the EAP against its role filler in the stored
 7440 tuple, and return this as the answer.

7441 To evaluate your system, download a set of three news articles on the same topic,
 7442 and write down five factoid questions that should be answerable from the arti-
 7443 cles. See if your system can answer these questions correctly. (If this problem is
 7444 assigned to an entire class, you can build a large-scale test set and compare various
 7445 approaches.)

¹⁶At the time of writing, the following systems are available: SENNA (<http://ronan.collobert.com/senna/>), Illinois Semantic Role Labeler (https://cogcomp.cs.illinois.edu/page/software_view/SRL), and mate-tools (<https://code.google.com/archive/p/mate-tools/>).

7446 Chapter 14

7447 Distributional and distributed 7448 semantics

7449 A recurring theme in natural language processing is the complexity of the mapping from
7450 words to meaning. In chapter 4, we saw that a single word form, like *bank*, can have mul-
7451 tiple meanings; conversely, a single meaning may be created by multiple surface forms,
7452 a lexical semantic relationship known as **synonymy**. Despite this complex mapping be-
7453 tween words and meaning, natural language processing systems usually rely on words
7454 as the basic unit of analysis. This is especially true in semantics: the logical and frame
7455 semantic methods from the previous two chapters rely on hand-crafted lexicons that map
7456 from words to semantic predicates. But how can we analyze texts that contain words
7457 that we haven't seen before? This chapter describes methods that learn representations
7458 of word meaning by analyzing unlabeled data, vastly improving the generalizability of
7459 natural language processing systems. The theory that makes it possible to acquire mean-
7460 ingful representations from unlabeled data is the **distributional hypothesis**.

7461 14.1 The distributional hypothesis

7462 Here's a word you may not know: *tezgüino* (the example is from Lin, 1998). If you do not
7463 know the meaning of *tezgüino*, then you are in the same situation as a natural language
7464 processing system when it encounters a word that did not appear in its training data.
7465 Now suppose you see that *tezgüino* is used in the following contexts:

- 7466 (14.1) A bottle of _____ is on the table.
- 7467 (14.2) Everybody likes _____.
- 7468 (14.3) Don't have _____ before you drive.
- 7469 (14.4) We make _____ out of corn.

	(14.1)	(14.2)	(14.3)	(14.4)	...
<i>tezgüino</i>	1	1	1	1	
<i>loud</i>	0	0	0	0	
<i>motor oil</i>	1	0	0	1	
<i>tortillas</i>	0	1	0	1	
<i>choices</i>	0	1	0	0	
<i>wine</i>	1	1	1	0	

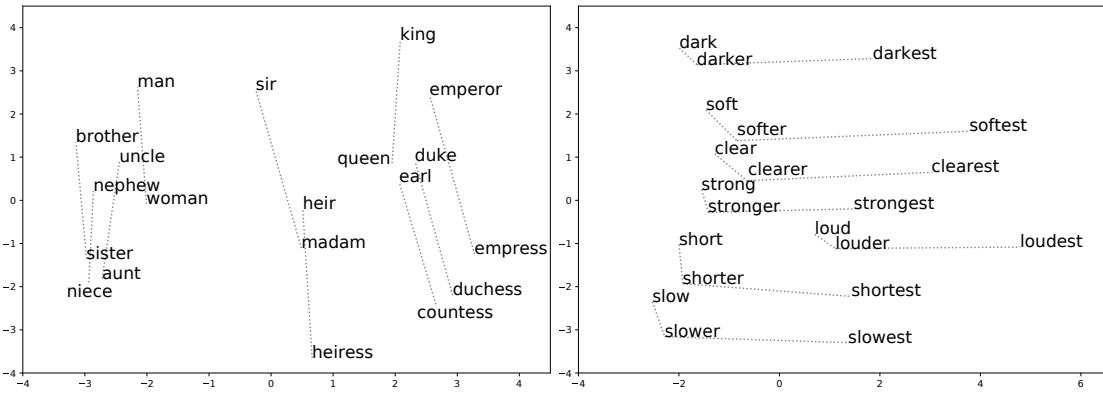
Table 14.1: Distributional statistics for *tezgüino* and five related terms

Figure 14.1: Lexical semantic relationships have regular linear structures in two dimensional projections of distributional statistics (Pennington et al., 2014).

7470 What other words fit into these contexts? How about: *loud*, *motor oil*, *tortillas*, *choices*,
 7471 *wine*? Each row of Table 14.1 is a vector that summarizes the contextual properties for
 7472 each word, with a value of one for contexts in which the word can appear, and a value of
 7473 zero for contexts in which it cannot. Based on these vectors, we can conclude: *wine* is very
 7474 similar to *tezgüino*; *motor oil* and *tortillas* are fairly similar to *tezgüino*; *loud* is completely
 7475 different.

7476 These vectors, which we will call **word representations**, describe the **distributional**
 7477 properties of each word. Does vector similarity imply semantic similarity? This is the **dis-**
 7478 **distributional hypothesis**, stated by Firth (1957) as: “You shall know a word by the company
 7479 it keeps.” The distributional hypothesis has stood the test of time: distributional statistics
 7480 are a core part of language technology today, because they make it possible to leverage
 7481 large amounts of unlabeled data to learn about rare words that do not appear in labeled
 7482 training data.

7483 Distributional statistics have a striking ability to capture lexical semantic relationships

such as analogies. Figure 14.1 shows two examples, based on two-dimensional projections of distributional **word embeddings**, discussed later in this chapter. In each case, word-pair relationships correspond to regular linear patterns in this two dimensional space. No labeled data about the nature of these relationships was required to identify this underlying structure.

Distributional semantics are computed from context statistics. **Distributed** semantics are a related but distinct idea: that meaning can be represented by numerical vectors rather than symbolic structures. Distributed representations are often estimated from distributional statistics, as in latent semantic analysis and WORD2VEC, described later in this chapter. However, distributed representations can also be learned in a supervised fashion from labeled data, as in the neural classification models encountered in chapter 3.

14.2 Design decisions for word representations

There are many approaches for computing word representations, but most can be distinguished on three main dimensions: the nature of the representation, the source of contextual information, and the estimation procedure.

14.2.1 Representation

Today, the dominant word representations are k -dimensional vectors of real numbers, known as **word embeddings**. (The name is due to the fact that each discrete word is embedded in a continuous vector space.) This representation dates back at least to the late 1980s (Deerwester et al., 1990), and is used in popular techniques such as WORD2VEC (Mikolov et al., 2013).

Word embeddings are well suited for neural networks, where they can be plugged in as inputs. They can also be applied in linear classifiers and structure prediction models (Turian et al., 2010), although it can be difficult to learn linear models that employ real-valued features (Kummerfeld et al., 2015). A popular alternative is bit-string representations, such as **Brown clusters** (§ 14.4), in which each word is represented by a variable-length sequence of zeros and ones (Brown et al., 1992).

Another representational question is whether to estimate one embedding per surface form (e.g., *bank*), or to estimate distinct embeddings for each word sense or synset. Intuitively, if word representations are to capture the meaning of individual words, then words with multiple meanings should have multiple embeddings. This can be achieved by integrating unsupervised clustering with word embedding estimation (Huang and Yates, 2012; Li and Jurafsky, 2015). However, Arora et al. (2018) argue that it is unnecessary to model distinct word senses explicitly, because the embeddings for each surface form are a linear combination of the embeddings of the underlying senses.

The moment one learns English, complications set in (Alfau, 1999)

Brown Clusters	$\{one\}$
WORD2VEC, $h = 2$	$\{moment, one, English, complications\}$
Structured WORD2VEC, $h = 2$	$\{(moment, -2), (one, -1), (English, +1), (complications, +2)\}$
Dependency contexts,	$\{(one, NSUBJ), (English, DOBJ), (moment, ACL^{-1})\}$

Table 14.2: Contexts for the word *learns*, according to various word representations. For dependency context, $(one, NSUBJ)$ means that there is a relation of type NSUBJ (nominal subject) *to* the word *one*, and $(moment, ACL^{-1})$ means that there is a relation of type ACL (adjectival clause) *from* the word *moment*.

7519 14.2.2 Context

7520 The distributional hypothesis says that word meaning is related to the “contexts” in which
 7521 the word appears, but context can be defined in many ways. In the *tezgiino* example, con-
 7522 texts are entire sentences, but in practice there are far too many sentences. At the oppo-
 7523 site extreme, the context could be defined as the immediately preceding word; this is the
 7524 context considered in Brown clusters. WORD2VEC takes an intermediate approach, using
 7525 local neighborhoods of words (e.g., $h = 5$) as contexts (Mikolov et al., 2013). Contexts
 7526 can also be much larger: for example, in **latent semantic analysis**, each word’s context
 7527 vector includes an entry per document, with a value of one if the word appears in the
 7528 document (Deerwester et al., 1990); in **explicit semantic analysis**, these documents are
 7529 Wikipedia pages (Gabrilovich and Markovitch, 2007).

7530 In structured WORD2VEC, context words are labeled by their position with respect to
 7531 the target word w_m (e.g., two words before, one word after), which makes the result-
 7532 ing word representations more sensitive to syntactic differences (Ling et al., 2015). An-
 7533 other way to incorporate syntax is to perform parsing as a preprocessing step, and then
 7534 form context vectors from the dependency edges (Levy and Goldberg, 2014) or predicate-
 7535 argument relations (Lin, 1998). The resulting context vectors for several of these methods
 7536 are shown in Table 14.2.

7537 The choice of context has a profound effect on the resulting representations, which
 7538 can be viewed in terms of word similarity. Applying latent semantic analysis (§ 14.3) to
 7539 contexts of size $h = 2$ and $h = 30$ yields the following nearest-neighbors for the word
 7540 *dog*.¹

- 7541 • ($h = 2$): *cat, horse, fox, pet, rabbit, pig, animal, mongrel, sheep, pigeon*

¹The example is from lecture slides by Marco Baroni, Alessandro Lenci, and Stefan Evert, who applied latent semantic analysis to the British National Corpus. You can find an online demo here: <http://clic.cimec.unitn.it/infomap-query/>

- 7542 • ($h = 30$): *kennel, puppy, pet, bitch, terrier, rottweiler, canine, cat, to bark, Alsatian*

7543 Which word list is better? Each word in the $h = 2$ list is an animal, reflecting the fact that
 7544 locally, the word *dog* tends to appear in the same contexts as other animal types (e.g., *pet*
 7545 *the dog, feed the dog*). In the $h = 30$ list, nearly everything is dog-related, including specific
 7546 breeds such as *rottweiler* and *Alsatian*. The list also includes words that are not animals
 7547 (*kennel*), and in one case (*to bark*), is not a noun at all. The 2-word context window is more
 7548 sensitive to syntax, while the 30-word window is more sensitive to topic.

7549 **14.2.3 Estimation**

7550 Word embeddings are estimated by optimizing some objective: the likelihood of a set of
 7551 unlabeled data (or a closely related quantity), or the reconstruction of a matrix of context
 7552 counts, similar to Table 14.1.

7553 **Maximum likelihood estimation** Likelihood-based optimization is derived from the
 7554 objective $\log p(\mathbf{w}; \mathbf{U})$, where $\mathbf{U} \in \mathbb{R}^{K \times V}$ is matrix of word embeddings, and $\mathbf{w} =$
 7555 $\{w_m\}_{m=1}^M$ is a corpus, represented as a list of M tokens. Recurrent neural network lan-
 7556 guage models (§ 6.3) optimize this objective directly, backpropagating to the input word
 7557 embeddings through the recurrent structure. However, state-of-the-art word embeddings
 7558 employ huge corpora with hundreds of billions of tokens, and recurrent architectures are
 7559 difficult to scale to such data. As a result, likelihood-based word embeddings are usually
 7560 based on simplified likelihoods or heuristic approximations.

Matrix factorization The matrix $\mathbf{C} = \{\text{count}(i, j)\}$ stores the co-occurrence counts of
 word i and context j . Word representations can be obtained by approximately factoring
 this matrix, so that $\text{count}(i, j)$ is approximated by a function of a word embedding \mathbf{u}_i and
 a context embedding \mathbf{v}_j . These embeddings can be obtained by minimizing the norm of
 the reconstruction error,

$$\min_{\mathbf{u}, \mathbf{v}} \|\mathbf{C} - \tilde{\mathbf{C}}(\mathbf{u}, \mathbf{v})\|_F, \quad [14.1]$$

7561 where $\tilde{\mathbf{C}}(\mathbf{u}, \mathbf{v})$ is the approximate reconstruction resulting from the embeddings \mathbf{u} and
 7562 \mathbf{v} , and $\|\mathbf{X}\|_F$ indicates the Frobenius norm, $\sum_{i,j} x_{i,j}^2$. Rather than factoring the matrix of
 7563 word-context counts directly, it is often helpful to transform these counts using information-
 7564 theoretic metrics such as **pointwise mutual information** (PMI), described in the next sec-
 7565 tion.

7566 **14.3 Latent semantic analysis**

Latent semantic analysis (LSA) is one of the oldest approaches to distributed semantics (Deerwester et al., 1990). It induces continuous vector representations of words by

factoring a matrix of word and context counts, using **truncated singular value decomposition** (SVD),

$$\min_{\mathbf{U} \in \mathbb{R}^{V \times K}, \mathbf{S} \in \mathbb{R}^{K \times K}, \mathbf{V} \in \mathbb{R}^{|\mathcal{C}| \times K}} \|\mathbf{C} - \mathbf{USV}^\top\|_F \quad [14.2]$$

$$\text{s.t. } \mathbf{U}^\top \mathbf{U} = \mathbb{I} \quad [14.3]$$

$$\mathbf{V}^\top \mathbf{V} = \mathbb{I} \quad [14.4]$$

$$\forall i \neq j, \mathbf{S}_{i,j} = 0, \quad [14.5]$$

where V is the size of the vocabulary, $|\mathcal{C}|$ is the number of contexts, and K is size of the resulting embeddings, which are set equal to the rows of the matrix \mathbf{U} . The matrix \mathbf{S} is constrained to be diagonal (these diagonal elements are called the singular values), and the columns of the product \mathbf{SV}^\top provide descriptions of the contexts. Each element $c_{i,j}$ is then reconstructed as a **bilinear product**,

$$c_{i,j} \approx \sum_{k=1}^K u_{i,k} s_k v_{j,k}. \quad [14.6]$$

The objective is to minimize the sum of squared approximation errors. The orthonormality constraints $\mathbf{U}^\top \mathbf{U} = \mathbf{V}^\top \mathbf{V} = \mathbb{I}$ ensure that all pairs of dimensions in \mathbf{U} and \mathbf{V} are uncorrelated, so that each dimension conveys unique information. Efficient implementations of truncated singular value decomposition are available in numerical computing packages such as SCIPY and MATLAB.²

Latent semantic analysis is most effective when the count matrix is transformed before the application of SVD. One such transformation is **pointwise mutual information** (PMI; Church and Hanks, 1990), which captures the degree of association between word i and context j ,

$$\text{PMI}(i, j) = \log \frac{\text{p}(i, j)}{\text{p}(i)\text{p}(j)} = \log \frac{\text{p}(i | j)\text{p}(j)}{\text{p}(i)\text{p}(j)} = \log \frac{\text{p}(i | j)}{\text{p}(i)} \quad [14.7]$$

$$= \log \text{count}(i, j) - \log \sum_{i'=1}^V \text{count}(i', j) \quad [14.8]$$

$$- \log \sum_{j' \in \mathcal{C}} \text{count}(i, j') + \log \sum_{i'=1}^V \sum_{j' \in \mathcal{C}} \text{count}(i', j'). \quad [14.9]$$

The pointwise mutual information can be viewed as the logarithm of the ratio of the conditional probability of word i in context j to the marginal probability of word i in all

²An important implementation detail is to represent \mathbf{C} as a **sparse matrix**, so that the storage cost is equal to the number of non-zero entries, rather than the size $V \times |\mathcal{C}|$.

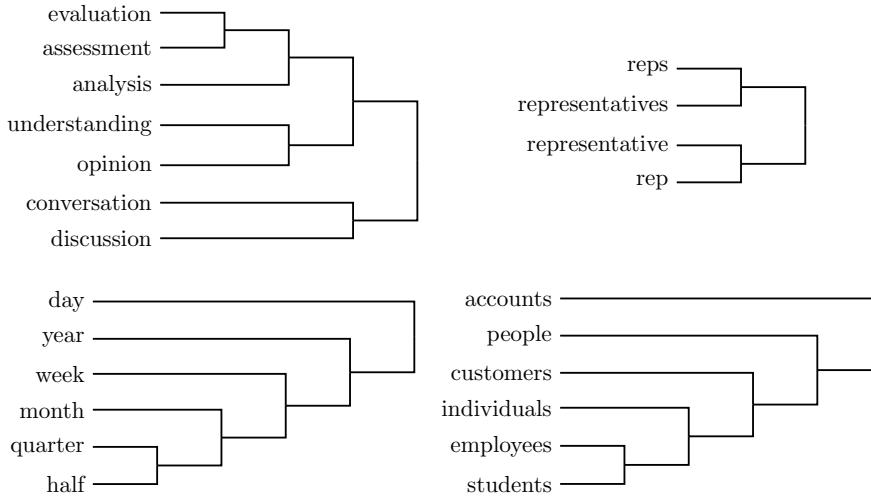


Figure 14.2: Subtrees produced by bottom-up Brown clustering on news text (Miller et al., 2004).

7579 contexts. When word i is statistically associated with context j , the ratio will be greater
 7580 than one, so $\text{PMI}(i, j) > 0$. The PMI transformation focuses latent semantic analysis on re-
 7581 constructing strong word-context associations, rather than on reconstructing large counts.

7582 The PMI is negative when a word and context occur together less often than if they
 7583 were independent, but such negative correlations are unreliable because counts of rare
 7584 events have high variance. Furthermore, the PMI is undefined when $\text{count}(i, j) = 0$. One
 7585 solution to these problems is to use the **Positive PMI** (PPMI),

$$\text{PPMI}(i, j) = \begin{cases} \text{PMI}(i, j), & p(i | j) > p(i) \\ 0, & \text{otherwise.} \end{cases} \quad [14.10]$$

7586 Bullinaria and Levy (2007) compare a range of matrix transformations for latent se-
 7587 mantic analysis, using a battery of tasks related to word meaning and word similarity
 7588 (for more on evaluation, see § 14.6). They find that PPMI-based latent semantic analysis
 7589 yields strong performance on a battery of tasks related to word meaning: for example,
 7590 PPMI-based LSA vectors can be used to solve multiple-choice word similarity questions
 7591 from the Test of English as a Foreign Language (TOEFL), obtaining 85% accuracy.

7592 14.4 Brown clusters

7593 Learning algorithms like perceptron and conditional random fields often perform better
 7594 with discrete feature vectors. A simple way to obtain discrete representations from distri-

bitstring	ten most frequent words
01111010 0111	<i>excited thankful grateful stoked pumped anxious hyped psyched exited geeked</i>
01111010 100	<i>talking talkin complaining talkn bitching tlkn tlkin bragging raving +k</i>
01111010 1010	<i>thinking thinkin dreaming worrying thinkn speakin reminiscing dreamin daydreaming fantasizing</i>
01111010 1011	<i>saying sayin suggesting stating sayn jokin talmbout implying insisting 5'2</i>
01111010 1100	<i>wonder dunno wondered duno donno dno dono wonda wounder dunnoe</i>
01111010 1101	<i>wondering wonders debating deciding pondering unsure wonderin debatin woundering wondern</i>
01111010 1110	<i>sure suree suuure suure sure- surre sures shuree</i>

Table 14.3: Fragment of a Brown clustering of Twitter data (Owoputi et al., 2013). Each row is a leaf in the tree, showing the ten most frequent words. This part of the tree emphasizes verbs of communicating and knowing, especially in the present participle. Each leaf node includes orthographic variants (*thinking*, *thinkin*, *thinkn*), semantically related terms (*excited*, *thankful*, *grateful*), and some outliers (*5'2*, *+k*). See http://www.cs.cmu.edu/~ark/TweetNLP/cluster_viewer.html for more.

7595 butional statistics is by clustering (§ 5.1.1), so that words in the same cluster have similar
 7596 distributional statistics. This can help in downstream tasks, by sharing features between
 7597 all words in the same cluster. However, there is an obvious tradeoff: if the number of clusters
 7598 is too small, the words in each cluster will not have much in common; if the number
 7599 of clusters is too large, then the learner will not see enough examples from each cluster to
 7600 generalize.

7601 A solution to this problem is **hierarchical clustering**: using the distributional statistics
 7602 to induce a tree-structured representation. Fragments of **Brown cluster** trees are shown in
 7603 Figure 14.2 and Table 14.3. Each word’s representation consists of a binary string describ-
 7604 ing a path through the tree: 0 for taking the left branch, and 1 for taking the right branch.
 7605 In the subtree in the upper right of the figure, the representation of the word *conversation*
 7606 is 10; the representation of the word *assessment* is 0001. Bitstring prefixes capture simi-
 7607 larity at varying levels of specificity, and it is common to use the first eight, twelve, sixteen,
 7608 and twenty bits as features in tasks such as named entity recognition (Miller et al., 2004)
 7609 and dependency parsing (Koo et al., 2008).

Hierarchical trees can be induced from a likelihood-based objective, using a discrete

latent variable $k_i \in \{1, 2, \dots, K\}$ to represent the cluster of word i :

$$\log p(\mathbf{w}; \mathbf{k}) \approx \sum_{m=1}^M \log p(w_m | w_{m-1}; \mathbf{k}) \quad [14.11]$$

$$\triangleq \sum_{m=1}^M \log p(w_m | k_{w_m}) + \log p(k_{w_m} | k_{w_{m-1}}). \quad [14.12]$$

- 7610 This is similar to a hidden Markov model, with the crucial difference that each word can
 7611 be emitted from only a single cluster: $\forall k \neq k_{w_m}, p(w_m | k) = 0$.

Using the objective in Equation 14.12, the Brown clustering tree can be constructed from the bottom up: begin with each word in its own cluster, and incrementally merge clusters until only a single cluster remains. At each step, we merge the pair of clusters such that the objective in Equation 14.12 is maximized. Although the objective seems to involve a sum over the entire corpus, the score for each merger can be computed from the cluster-to-cluster co-occurrence counts. These counts can be updated incrementally as the clustering proceeds. The optimal merge at each step can be shown to maximize the **average mutual information**,

$$I(\mathbf{k}) = \sum_{k_1=1}^K \sum_{k_2=1}^K p(k_1, k_2) \times \text{PMI}(k_1, k_2) \quad [14.13]$$

$$p(k_1, k_2) = \frac{\text{count}(k_1, k_2)}{\sum_{k_1'=1}^K \sum_{k_2'=1}^K \text{count}(k_1', k_2')},$$

- 7612 where $p(k_1, k_2)$ is the joint probability of a bigram involving a word in cluster k_1 followed
 7613 by a word in k_2 . This probability and the PMI are both computed from the co-occurrence
 7614 counts between clusters. After each merger, the co-occurrence vectors for the merged
 7615 clusters are simply added up, so that the next optimal merger can be found efficiently.

- 7616 This bottom-up procedure requires iterating over the entire vocabulary, and evaluating
 7617 K_t^2 possible mergers at each step, where K_t is the current number of clusters at step t
 7618 of the algorithm. Furthermore, computing the score for each merger involves a sum over
 7619 K_t^2 clusters. The maximum number of clusters is $K_0 = V$, which occurs when every word
 7620 is in its own cluster at the beginning of the algorithm. The time complexity is thus $\mathcal{O}(V^5)$.

- 7621 To avoid this complexity, practical implementations use a heuristic approximation
 7622 called **exchange clustering**. The K most common words are placed in clusters of their
 7623 own at the beginning of the process. We then consider the next most common word, and
 7624 merge it with one of the existing clusters. This continues until the entire vocabulary has
 7625 been incorporated, at which point the K clusters are merged down to a single cluster,
 7626 forming a tree. The algorithm never considers more than $K + 1$ clusters at any step, and
 7627 the complexity is $\mathcal{O}(VK + V \log V)$, with the second term representing the cost of sorting

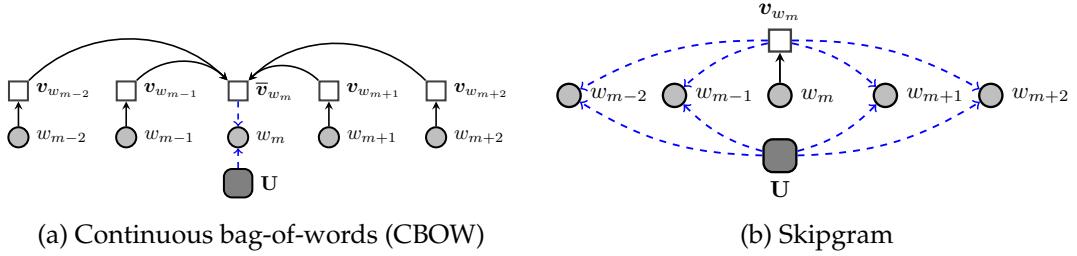


Figure 14.3: The CBOW and skipgram variants of WORD2VEC. The parameter \mathbf{U} is the matrix of word embeddings, and each v_m is the context embedding for word w_m .

7628 the words at the beginning of the algorithm. For more details on the algorithm, see Liang
 7629 (2005).

7630 14.5 Neural word embeddings

7631 Neural word embeddings combine aspects of the previous two methods: like latent se-
 7632 mantic analysis, they are a continuous vector representation; like Brown clusters, they are
 7633 trained from a likelihood-based objective. Let the vector u_i represent the K -dimensional
 7634 **embedding** for word i , and let v_j represent the K -dimensional embedding for context
 7635 j . The inner product $u_i \cdot v_j$ represents the compatibility between word i and context j .
 7636 By incorporating this inner product into an approximation to the log-likelihood of a cor-
 7637 pus, it is possible to estimate both parameters by backpropagation. WORD2VEC (Mikolov
 7638 et al., 2013) includes two such approximations: continuous bag-of-words (CBOW) and
 7639 skipgrams.

7640 14.5.1 Continuous bag-of-words (CBOW)

7641 In recurrent neural network language models, each word w_m is conditioned on a recurrently-
 7642 updated state vector, which is based on word representations going all the way back to the
 7643 beginning of the text. The **continuous bag-of-words (CBOW)** model is a simplification:
 7644 the local context is computed as an average of embeddings for words in the immediate
 7645 neighborhood $m - h, m - h + 1, \dots, m + h - 1, m + h$,

$$\bar{v}_m = \frac{1}{2h} \sum_{n=1}^h v_{w_{m+n}} + v_{w_{m-n}}. \quad [14.14]$$

7646 Thus, CBOW is a bag-of-words model, because the order of the context words does not
 7647 matter; it is continuous, because rather than conditioning on the words themselves, we
 7648 condition on a continuous vector constructed from the word embeddings. The parameter
 7649 h determines the neighborhood size, which Mikolov et al. (2013) set to $h = 4$.

The CBOW model optimizes an approximation to the corpus log-likelihood,

$$\log p(\mathbf{w}) \approx \sum_{m=1}^M \log p(w_m | w_{m-h}, w_{m-h+1}, \dots, w_{m+h-1}, w_{m+h}) \quad [14.15]$$

$$= \sum_{m=1}^M \log \frac{\exp(\mathbf{u}_{w_m} \cdot \bar{\mathbf{v}}_m)}{\sum_{j=1}^V \exp(\mathbf{u}_j \cdot \bar{\mathbf{v}}_m)} \quad [14.16]$$

$$= \sum_{m=1}^M \mathbf{u}_{w_m} \cdot \bar{\mathbf{v}}_m - \log \sum_{j=1}^V \exp(\mathbf{u}_j \cdot \bar{\mathbf{v}}_m). \quad [14.17]$$

7650 14.5.2 Skipgrams

In the CBOW model, words are predicted from their context. In the **skipgram** model, the context is predicted from the word, yielding the objective:

$$\log p(\mathbf{w}) \approx \sum_{m=1}^M \sum_{n=1}^{h_m} \log p(w_{m-n} | w_m) + \log p(w_{m+n} | w_m) \quad [14.18]$$

$$= \sum_{m=1}^M \sum_{n=1}^{h_m} \log \frac{\exp(\mathbf{u}_{w_{m-n}} \cdot \mathbf{v}_{w_m})}{\sum_{j=1}^V \exp(\mathbf{u}_j \cdot \mathbf{v}_{w_m})} + \log \frac{\exp(\mathbf{u}_{w_{m+n}} \cdot \mathbf{v}_{w_m})}{\sum_{j=1}^V \exp(\mathbf{u}_j \cdot \mathbf{v}_{w_m})} \quad [14.19]$$

$$= \sum_{m=1}^M \sum_{n=1}^{h_m} \mathbf{u}_{w_{m-n}} \cdot \mathbf{v}_{w_m} + \mathbf{u}_{w_{m+n}} \cdot \mathbf{v}_{w_m} - 2 \log \sum_{j=1}^V \exp(\mathbf{u}_j \cdot \mathbf{v}_{w_m}). \quad [14.20]$$

7651 In the skipgram approximation, each word is generated multiple times; each time it is con-
 7652 ditioned only on a single word. This makes it possible to avoid averaging the word vec-
 7653 tors, as in the CBOW model. The local neighborhood size h_m is randomly sampled from
 7654 a uniform categorical distribution over the range $\{1, 2, \dots, h_{\max}\}$; Mikolov et al. (2013) set
 7655 $h_{\max} = 10$. Because the neighborhood grows outward with h , this approach has the effect
 7656 of weighting near neighbors more than distant ones. Skipgram performs better on most
 7657 evaluations than CBOW (see § 14.6 for details of how to evaluate word representations),
 7658 but CBOW is faster to train (Mikolov et al., 2013).

7659 14.5.3 Computational complexity

7660 The WORD2VEC models can be viewed as an efficient alternative to recurrent neural net-
 7661 work language models, which involve a recurrent state update whose time complexity
 7662 is quadratic in the size of the recurrent state vector. CBOW and skipgram avoid this
 7663 computation, and incur only a linear time complexity in the size of the word and con-
 7664 text representations. However, all three models compute a normalized probability over
 7665 word tokens; a naïve implementation of this probability requires summing over the entire

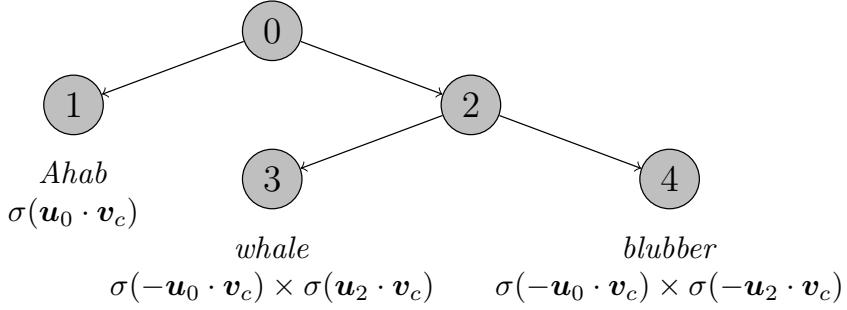


Figure 14.4: A fragment of a hierarchical softmax tree. The probability of each word is computed as a product of probabilities of local branching decisions in the tree.

vocabulary. The time complexity of this sum is $\mathcal{O}(V \times K)$, which dominates all other computational costs. There are two solutions: **hierarchical softmax**, a tree-based computation that reduces the cost to a logarithm of the size of the vocabulary; and **negative sampling**, an approximation that eliminates the dependence on vocabulary size. Both methods are also applicable to RNN language models.

7671 Hierarchical softmax

In Brown clustering, the vocabulary is organized into a binary tree. Mnih and Hinton (2008) show that the normalized probability over words in the vocabulary can be reparametrized as a probability over paths through such a tree. This hierarchical softmax probability is computed as a product of binary decisions over whether to move left or right through the tree, with each binary decision represented as a sigmoid function of the inner product between the context embedding v_c and an output embedding associated with the node u_n ,

$$\Pr(\text{left at } n \mid c) = \sigma(u_n \cdot v_c) \quad [14.21]$$

$$\Pr(\text{right at } n \mid c) = 1 - \sigma(u_n \cdot v_c) = \sigma(-u_n \cdot v_c), \quad [14.22]$$

where σ refers to the sigmoid function, $\sigma(x) = \frac{1}{1+\exp(-x)}$. The range of the sigmoid is the interval $(0, 1)$, and $1 - \sigma(x) = \sigma(-x)$.

As shown in Figure 14.4, the probability of generating each word is redefined as the product of the probabilities across its path. The sum of all such path probabilities is guaranteed to be one, for any context vector $v_c \in \mathbb{R}^K$. In a balanced binary tree, the depth is logarithmic in the number of leaf nodes, and thus the number of multiplications is equal to $\mathcal{O}(\log V)$. The number of non-leaf nodes is equal to $\mathcal{O}(2V - 1)$, so the number of parameters to be estimated increases by only a small multiple. The tree can be constructed using an incremental clustering procedure similar to hierarchical Brown clusters (Mnih

7681 and Hinton, 2008), or by using the Huffman (1952) encoding algorithm for lossless com-
 7682 pression.

7683 **Negative sampling**

Likelihood-based methods are computationally intensive because each probability must be normalized over the vocabulary. These probabilities are based on scores for each word in each context, and it is possible to design an alternative objective that is based on these scores more directly: we seek word embeddings that maximize the score for the word that was really observed in each context, while minimizing the scores for a set of randomly selected **negative samples**:

$$\psi(i, j) = \log \sigma(\mathbf{u}_i \cdot \mathbf{v}_j) + \sum_{i' \in \mathcal{W}_{\text{neg}}} \log(1 - \sigma(\mathbf{u}_{i'} \cdot \mathbf{v}_j)), \quad [14.23]$$

7684 where $\psi(i, j)$ is the score for word i in context j , and \mathcal{W}_{neg} is the set of negative samples.
 7685 The objective is to maximize the sum over the corpus, $\sum_{m=1}^M \psi(w_m, c_m)$, where w_m is
 7686 token m and c_m is the associated context.

7687 The set of negative samples \mathcal{W}_{neg} is obtained by sampling from a unigram language
 7688 model. Mikolov et al. (2013) construct this unigram language model by exponentiating
 7689 the empirical word probabilities, setting $\hat{p}(i) \propto (\text{count}(i))^{\frac{3}{4}}$. This has the effect of redis-
 7690 tributing probability mass from common to rare words. The number of negative samples
 7691 increases the time complexity of training by a constant factor. Mikolov et al. (2013) report
 7692 that 5-20 negative samples works for small training sets, and that two to five samples
 7693 suffice for larger corpora.

7694 **14.5.4 Word embeddings as matrix factorization**

7695 The negative sampling objective in Equation 14.23 can be justified as an efficient approx-
 7696 imation to the log-likelihood, but it is also closely linked to the matrix factorization ob-
 7697 jective employed in latent semantic analysis. For a matrix of word-context pairs in which
 7698 all counts are non-zero, negative sampling is equivalent to factorization of the matrix M ,
 7699 where $M_{ij} = \text{PMI}(i, j) - \log k$: each cell in the matrix is equal to the pointwise mutual
 7700 information of the word and context, shifted by $\log k$, with k equal to the number of neg-
 7701 ative samples (Levy and Goldberg, 2014). For word-context pairs that are not observed in
 7702 the data, the pointwise mutual information is $-\infty$, but this can be addressed by consid-
 7703 ering only PMI values that are greater than $\log k$, resulting in a matrix of **shifted positive**
 7704 **pointwise mutual information**,

$$M_{ij} = \max(0, \text{PMI}(i, j) - \log k). \quad [14.24]$$

7705 Word embeddings are obtained by factoring this matrix with truncated singular value
 7706 decomposition.

word 1	word 2	similarity
<i>love</i>	<i>sex</i>	6.77
<i>stock</i>	<i>jaguar</i>	0.92
<i>money</i>	<i>cash</i>	9.15
<i>development</i>	<i>issue</i>	3.97
<i>lad</i>	<i>brother</i>	4.46

Table 14.4: Subset of the WS-353 (Finkelstein et al., 2002) dataset of word similarity ratings (examples from Faruqui et al. (2016)).

GloVe (“global vectors”) are a closely related approach (Pennington et al., 2014), in which the matrix to be factored is constructed from log co-occurrence counts, $M_{ij} = \log \text{count}(i, j)$. The word embeddings are estimated by minimizing the sum of squares,

$$\begin{aligned} \min_{\mathbf{u}, \mathbf{v}, b, \tilde{b}} \quad & \sum_{j=1}^V \sum_{j \in \mathcal{C}} f(M_{ij}) \left(\widehat{\log M_{ij}} - \log M_{ij} \right)^2 \\ \text{s.t.} \quad & \widehat{\log M_{ij}} = \mathbf{u}_i \cdot \mathbf{v}_j + b_i + \tilde{b}_j, \end{aligned} \quad [14.25]$$

7707 where b_i and \tilde{b}_j are offsets for word i and context j , which are estimated jointly with the
 7708 embeddings \mathbf{u} and \mathbf{v} . The weighting function $f(M_{ij})$ is set to be zero at $M_{ij} = 0$, thus
 7709 avoiding the problem of taking the logarithm of zero counts; it saturates at $M_{ij} = m_{\max}$,
 7710 thus avoiding the problem of overcounting common word-context pairs. This heuristic
 7711 turns out to be critical to the method’s performance.

7712 The time complexity of sparse matrix reconstruction is determined by the number of
 7713 non-zero word-context counts. Pennington et al. (2014) show that this number grows
 7714 sublinearly with the size of the dataset: roughly $\mathcal{O}(N^{0.8})$ for typical English corpora. In
 7715 contrast, the time complexity of WORD2VEC is linear in the corpus size. Computing the co-
 7716 occurrence counts also requires linear time in the size of the corpus, but this operation can
 7717 easily be parallelized using MapReduce-style algorithms (Dean and Ghemawat, 2008).

7718 14.6 Evaluating word embeddings

7719 Distributed word representations can be evaluated in two main ways. **Intrinsic** evalua-
 7720 tions test whether the representations cohere with our intuitions about word meaning.
 7721 **Extrinsic** evaluations test whether they are useful for downstream tasks, such as sequence
 7722 labeling.

7723 **14.6.1 Intrinsic evaluations**

7724 A basic question for word embeddings is whether the similarity of words i and j is re-
 7725 flected in the similarity of the vectors \mathbf{u}_i and \mathbf{u}_j . **Cosine similarity** is typically used to
 7726 compare two word embeddings,

$$\cos(\mathbf{u}_i, \mathbf{u}_j) = \frac{\mathbf{u}_i \cdot \mathbf{u}_j}{\|\mathbf{u}_i\|_2 \times \|\mathbf{u}_j\|_2}. \quad [14.26]$$

7727 For any embedding method, we can evaluate whether the cosine similarity of word em-
 7728 beddings is correlated with human judgments of word similarity. The WS-353 dataset (Finkel-
 7729 stein et al., 2002) includes similarity scores for 353 word pairs (Table 14.4). To test the
 7730 accuracy of embeddings for rare and morphologically complex words, Luong et al. (2013)
 7731 introduce a dataset of “rare words.” Outside of English, word similarity resources are lim-
 7732 ited, mainly consisting of translations of WS-353 and the related SimLex-999 dataset (Hill
 7733 et al., 2015).

7734 Word analogies (e.g., *king:queen :: man:woman*) have also been used to evaluate word
 7735 embeddings (Mikolov et al., 2013). In this evaluation, the system is provided with the first
 7736 three parts of the analogy ($i_1 : j_1 :: i_2 : ?$), and the final element is predicted by finding the
 7737 word embedding most similar to $\mathbf{u}_{i_1} - \mathbf{u}_{j_1} + \mathbf{u}_{i_2}$. Another evaluation tests whether word
 7738 embeddings are related to broad lexical semantic categories called **supersenses** (Ciaramita
 7739 and Johnson, 2003): verbs of motion, nouns that describe animals, nouns that describe
 7740 body parts, and so on. These supersenses are annotated for English synsets in Word-
 7741 Net (Fellbaum, 2010). This evaluation is implemented in the QVEC metric, which tests
 7742 whether the matrix of supersenses can be reconstructed from the matrix of word embed-
 7743 dings (Tsvetkov et al., 2015).

7744 Levy et al. (2015) compared several dense word representations for English — includ-
 7745 ing latent semantic analysis, WORD2VEC, and GloVe — using six word similarity metrics
 7746 and two analogy tasks. None of the embeddings outperformed the others on every task,
 7747 but skipgrams were the most broadly competitive. Hyperparameter tuning played a key
 7748 role: any method will perform badly if the wrong hyperparameters are used. Relevant
 7749 hyperparameters include the embedding size, as well as algorithm-specific details such
 7750 as the neighborhood size and the number of negative samples.

7751 **14.6.2 Extrinsic evaluations**

7752 Word representations contribute to downstream tasks like sequence labeling and docu-
 7753 ment classification by enabling generalization across words. The use of distributed repre-
 7754 sentations as features is a form of **semi-supervised learning**, in which performance on a
 7755 supervised learning problem is augmented by learning distributed representations from
 7756 unlabeled data (Miller et al., 2004; Koo et al., 2008; Turian et al., 2010). These **pre-trained**
 7757 **word representations** can be used as features in a linear prediction model, or as the input

layer in a neural network, such as a Bi-LSTM tagging model (§ 7.6). Word representations can be evaluated by the performance of the downstream systems that consume them: for example, GloVe embeddings are convincingly better than Latent Semantic Analysis as features in the downstream task of named entity recognition (Pennington et al., 2014). Unfortunately, extrinsic and intrinsic evaluations do not always point in the same direction, and the best word representations for one downstream task may perform poorly on another task (Schnabel et al., 2015).

When word representations are updated from labeled data in the downstream task, they are said to be **fine-tuned**. When labeled data is plentiful, pre-training may be unnecessary; when labeled data is scarce, fine-tuning may lead to overfitting. Various combinations of pre-training and fine-tuning can be employed. Pre-trained embeddings can be used as initialization before fine-tuning, and this can substantially improve performance (Lample et al., 2016). Alternatively, both fine-tuned and pre-trained embeddings can be used as inputs in a single model (Kim, 2014).

In semi-supervised scenarios, pretrained word embeddings can be replaced by “contextualized” word representations (Peters et al., 2018). These contextualized representations are set to the hidden states of a deep bi-directional LSTM, which is trained as a bi-directional language model, motivating the name **ELMo (embeddings from language models)**. By running the language model, we obtain contextualized word representations, which can then be used as the base layer in a supervised neural network for any task. This approach yields significant gains over pretrained word embeddings on several tasks, presumably because the contextualized embeddings use unlabeled data to learn how to integrate linguistic context into the base layer of the supervised neural network.

14.6.3 Fairness and bias

Figure 14.1 shows how word embeddings can capture analogies such as *man:woman :: king:queen*. While *king* and *queen* are gender-specific by definition, other professions or titles are associated with genders and other groups merely by statistical tendency. This statistical tendency may be a fact about the world (e.g., professional baseball players are usually men), or a fact about the text corpus (e.g., there are professional basketball leagues for both women and men, but the men’s basketball is written about far more often).

There is now considerable evidence that word embeddings do indeed encode such biases. Bolukbasi et al. (2016) show that the words most aligned with the vector difference *she – he* are stereotypically female professions *homemaker, nurse, receptionist*; in the other direction are *maestro, skipper, protege*. Caliskan et al. (2017) systematize this observation by showing that biases in word embeddings align with well-validated gender stereotypes. Garg et al. (2018) extend these results to ethnic stereotypes of Asian Americans, and provide a historical perspective on how stereotypes evolve over 100 years of text data.

Because word embeddings are the input layer for many other natural language pro-

cessing systems, these findings highlight the risk that natural language processing will replicate and amplify biases in the world, as well as in text. If, for example, word embeddings encode the belief that women are as unlikely to be computer programmers as they are to be nephews, then software is unlikely to successfully parse, translate, index, and extract those cases in which women do indeed program computers. For example, in such cases, contemporary NLP systems often fail to properly resolve pronoun references (Rudinger et al., 2018; Zhao et al., 2018). (The task of pronoun resolution is described in depth in chapter 15.) Such biases can have profound consequences: for example, search engines are more likely to yield personalized advertisements for public arrest records when queried with names that are statistically associated with African Americans (Sweeney, 2013). There is now an active research literature on “debiasing” machine learning and natural language processing, as evidenced by the growth of annual meetings such as Fairness, Accountability, and Transparency in Machine Learning (FAT/ML). However, given that the ultimate source of these biases is the text itself, it may be too much to hope for a purely algorithmic solution. There is no substitute for critical thought about the inputs to natural language processing systems – and the uses of their outputs.

14.7 Distributed representations beyond distributional statistics

Distributional word representations can be estimated from huge unlabeled datasets, thereby covering many words that do not appear in labeled data: for example, GloVe embeddings are estimated from 800 billion tokens of web data,³ while the largest labeled datasets for NLP tasks are on the order of millions of tokens. Nonetheless, even a dataset of hundreds of billions of tokens will not cover every word that may be encountered in the future. Furthermore, many words will appear only a few times, making their embeddings unreliable. Many languages exceed English in morphological complexity, and thus have lower token-to-type ratios. When this problem is coupled with small training corpora, it becomes especially important to leverage other sources of information beyond distributional statistics.

14.7.1 Word-internal structure

One solution is to incorporate word-internal structure into word embeddings. Purely distributional approaches consider words as atomic units, but in fact, many words have internal structure, so that their meaning can be **composed** from the representations of sub-word units. Consider the following terms, all of which are missing from Google’s pre-trained WORD2VEC embeddings:⁴

³<http://commoncrawl.org/>

⁴<https://code.google.com/archive/p/word2vec/>, accessed September 20, 2017

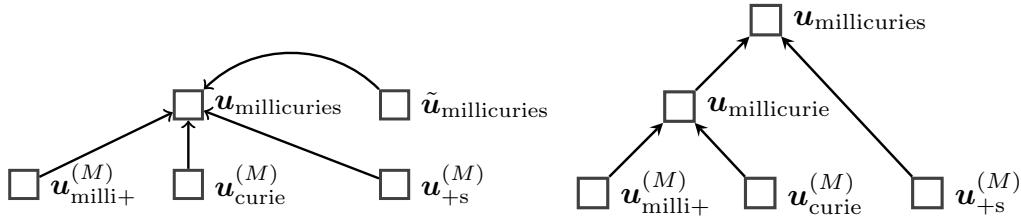


Figure 14.5: Two architectures for building word embeddings from subword units. On the left, morpheme embeddings $u^{(m)}$ are combined by addition with the non-compositional word embedding \tilde{u} (Botha and Blunsom, 2014). On the right, morpheme embeddings are combined in a recursive neural network (Luong et al., 2013).

7829 ***millicuries*** This word has **morphological** structure (see § 9.1.2 for more on morphology):
 7830 the prefix *milli-* indicates an amount, and the suffix *-s* indicates a plural. (A *millicurie*
 7831 is an unit of radioactivity.)

7832 ***caesium*** This word is a single morpheme, but the characters *-ium* are often associated
 7833 with chemical elements. (*Caesium* is the British spelling of a chemical element,
 7834 spelled *cesium* in American English.)

7835 ***IAEA*** This term is an acronym, as suggested by the use of capitalization. The prefix *I-* fre-
 7836 quently refers to international organizations, and the suffix *-A* often refers to agen-
 7837 cies or associations. (*IAEA* is the International Atomic Energy Agency.)

7838 ***Zhezhan*** This term is in title case, suggesting the name of a person or place, and the
 7839 character bigram *zh* indicates that it is likely a transliteration. (*Zhezhan* is a mining
 7840 facility in Kazakhstan.)

7841 How can word-internal structure be incorporated into word representations? One
 7842 approach is to construct word representations from embeddings of the characters or mor-
 7843 phemes. For example, if word i has morphological segments \mathcal{M}_i , then its embedding can
 7844 be constructed by addition (Botha and Blunsom, 2014),

$$\mathbf{u}_i = \tilde{\mathbf{u}}_i + \sum_{j \in \mathcal{M}_i} \mathbf{u}_j^{(M)}, \quad [14.27]$$

7845 where $\mathbf{u}_m^{(M)}$ is a morpheme embedding and $\tilde{\mathbf{u}}_i$ is a non-compositional embedding of the
 7846 whole word, which is an additional free parameter of the model (Figure 14.5, left side).
 7847 All embeddings are estimated from a **log-bilinear language model** (Mnih and Hinton,
 7848 2007), which is similar to the CBOW model (§ 14.5), but includes only contextual informa-
 7849 tion from preceding words. The morphological segments are obtained using an unsuper-
 7850 vised segmenter (Creutz and Lagus, 2007). For words that do not appear in the training

7851 data, the embedding can be constructed directly from the morphemes, assuming that each
 7852 morpheme appears in some other word in the training data. The free parameter \tilde{u} adds
 7853 flexibility: words with similar morphemes are encouraged to have similar embeddings,
 7854 but this parameter makes it possible for them to be different.

7855 Word-internal structure can be incorporated into word representations in various other
 7856 ways. Here are some of the main parameters.

7857 **Subword units.** Examples like *IAEA* and *Zhezhgan* are not based on morphological com-
 7858 position, and a morphological segmenter is unlikely to identify meaningful sub-
 7859 word units for these terms. Rather than using morphemes for subword embeddings,
 7860 one can use characters (Santos and Zadrozny, 2014; Ling et al., 2015; Kim et al., 2016),
 7861 character n -grams (Wieting et al., 2016a; Bojanowski et al., 2017), and **byte-pair en-**
 7862 **codings**, a compression technique which captures frequent substrings (Gage, 1994;
 7863 Sennrich et al., 2016).

7864 **Composition.** Combining the subword embeddings by addition does not differentiate
 7865 between orderings, nor does it identify any particular morpheme as the root. A
 7866 range of more flexible compositional models have been considered, including re-
 7867 currence (Ling et al., 2015), convolution (Santos and Zadrozny, 2014; Kim et al.,
 7868 2016), and **recursive neural networks** (Luong et al., 2013), in which representa-
 7869 tions of progressively larger units are constructed over a morphological parse, e.g.
 7870 $((milli+curie)+s)$, $((in+flam)+able)$, $(in+(vis+ible))$. A recursive embedding model is
 7871 shown in the right panel of Figure 14.5.

7872 **Estimation.** Estimating subword embeddings from a full dataset is computationally ex-
 7873 pensive. An alternative approach is to train a subword model to match pre-trained
 7874 word embeddings (Cotterell et al., 2016; Pinter et al., 2017). To train such a model, it
 7875 is only necessary to iterate over the vocabulary, and the not the corpus.

7876 14.7.2 Lexical semantic resources

Resources such as WordNet provide another source of information about word meaning; if we know that *caesium* is a synonym of *cesium*, or that a *millicurie* is a type of *measurement unit*, then this should help to provide embeddings for the unknown words, and to smooth embeddings of rare words. One way to do this is to **retrofit** pre-trained word embeddings across a network of lexical semantic relationships (Faruqui et al., 2015) by minimizing the following objective,

$$\min_{\mathbf{U}} \sum_{j=1}^V \|\mathbf{u}_i - \hat{\mathbf{u}}_i\|_2 + \sum_{(i,j) \in \mathcal{L}} \beta_{ij} \|\mathbf{u}_i - \mathbf{u}_j\|_2, \quad [14.28]$$

7877 where \hat{u}_i is the pretrained embedding of word i , and $\mathcal{L} = \{(i, j)\}$ is a lexicon of word
 7878 relations. The hyperparameter β_{ij} controls the importance of adjacent words having
 7879 similar embeddings; Faruqui et al. (2015) set it to the inverse of the degree of word i ,
 7880 $\beta_{ij} = |\{j : (i, j) \in \mathcal{L}\}|^{-1}$. Retrofitting improves performance on a range of intrinsic evalua-
 7881 tions, and gives small improvements on an extrinsic document classification task.

7882 14.8 Distributed representations of multiword units

7883 Can distributed representations extend to phrases, sentences, paragraphs, and beyond?
 7884 Before exploring this possibility, recall the distinction between distributed and distri-
 7885 butional representations. Neural embeddings such as WORD2VEC are both distributed
 7886 (vector-based) and distributional (derived from counts of words in context). As we con-
 7887 sider larger units of text, the counts decrease: in the limit, a multi-paragraph span of text
 7888 would never appear twice, except by plagiarism. Thus, the meaning of a large span of
 7889 text cannot be determined from distributional statistics alone; it must be computed com-
 7890 positionally from smaller spans. But these considerations are orthogonal to the question
 7891 of whether distributed representations — dense numerical vectors — are sufficiently ex-
 7892 pressive to capture the meaning of phrases, sentences, and paragraphs.

7893 14.8.1 Purely distributional methods

7894 Some multiword phrases are non-compositional: the meaning of such phrases is not de-
 7895 rived from the meaning of the individual words using typical compositional semantics.
 7896 This includes proper nouns like *San Francisco* as well as idiomatic expressions like *kick*
 7897 *the bucket* (Baldwin and Kim, 2010). For these cases, purely distributional approaches
 7898 can work. A simple approach is to identify multiword units that appear together fre-
 7899 quently, and then treat these units as words, learning embeddings using a technique such
 7900 as WORD2VEC.

7901 The problem of identifying multiword units is sometimes called **collocation extrac-**
 7902 **tion.** A good collocation has high **pointwise mutual information** (PMI), $\log p(w_t =$
 7903 $i | w_{t-1} = j) - \log p(w_t = i)$. For example, *Naïve Bayes* is a good collocation because
 7904 $p(w_t = Bayes | w_{t-1} = naïve)$ is much larger than $p(w_t = Bayes)$. Multiword collocation
 7905 can be performed by greedily extracting and grouping the collocations with the maxi-
 7906 mum PMI: for example, *mutual information* might first be extracted as a collocation and
 7907 grouped into a single word type *mutual_information*; then *pointwise mutual_information* can
 7908 be extracted later. After identifying such units, they can be treated as words when esti-
 7909 mating skipgram embeddings. Mikolov et al. (2013) show that the resulting embeddings
 7910 perform reasonably on a task of solving phrasal analogies, e.g. *New York : New York Times*
 7911 $:: Baltimore : Baltimore Sun$.

this was the only way
 it was the only way
 it was her turn to blink
 it was hard to tell
 it was time to move on
 he had to do it again
 they all looked at each other
 they all turned to look back
 they both turned to face him
they both turned and walked away

Figure 14.6: By interpolating between the distributed representations of two sentences (in bold), it is possible to generate grammatical sentences that combine aspects of both (Bowman et al., 2016)

7912 **14.8.2 Distributional-compositional hybrids**

7913 To move beyond short multiword phrases, composition is necessary. A simple but sur-
 7914 prisingly powerful approach is to represent a sentence with the average of its word em-
 7915 beddings (Mitchell and Lapata, 2010). This can be considered a hybrid of the distribu-
 7916 tional and compositional approaches to semantics: the word embeddings are computed
 7917 distributionally, and then the sentence representation is computed by composition.

7918 The WORD2VEC approach can be stretched considerably further, embedding entire
 7919 sentences using a model similar to skipgrams, in the “skip-thought” model of Kiros et al.
 7920 (2015). Each sentence is *encoded* into a vector using a recurrent neural network: the encod-
 7921 ing of sentence t is set to the RNN hidden state at its final token, $h_{M_t}^{(t)}$. This vector is then
 7922 a parameter in a *decoder* model that is used to generate the previous and subsequent sen-
 7923 tences: the decoder is another recurrent neural network, which takes the encoding of the
 7924 neighboring sentence as an additional parameter in its recurrent update. (This **encoder-**
 7925 **decoder model** is discussed at length in chapter 18.) The encoder and decoder are trained
 7926 simultaneously from a likelihood-based objective, and the trained encoder can be used to
 7927 compute a distributed representation of any sentence. Skip-thought can also be viewed
 7928 as a hybrid of distributional and compositional approaches: the vector representation of
 7929 each sentence is computed compositionally from the representations of the individual
 7930 words, but the training objective is distributional, based on sentence co-occurrence across
 7931 a corpus.

7932 **Autoencoders** are a variant of encoder-decoder models in which the decoder is trained
 7933 to produce the same text that was originally encoded, using only the distributed encod-
 7934 ing vector (Li et al., 2015). The encoding acts as a bottleneck, so that generalization is
 7935 necessary if the model is to successfully fit the training data. In **denoising autoencoders**,

7936 the input is a corrupted version of the original sentence, and the auto-encoder must re-
 7937 construct the uncorrupted original (Vincent et al., 2010; Hill et al., 2016). By interpolating
 7938 between distributed representations of two sentences, $\alpha \mathbf{u}_i + (1 - \alpha) \mathbf{u}_j$, it is possible to gen-
 7939 erate sentences that combine aspects of the two inputs, as shown in Figure 14.6 (Bowman
 7940 et al., 2016).

7941 Autoencoders can also be applied to longer texts, such as paragraphs and documents.
 7942 This enables applications such as **question answering**, which can be performed by match-
 7943 ing the encoding of the question with encodings of candidate answers (Miao et al., 2016).

7944 14.8.3 Supervised compositional methods

7945 Given a supervision signal, such as a label describing the sentiment or meaning of a sen-
 7946 tence, a wide range of compositional methods can be applied to compute a distributed
 7947 representation that then predicts the label. The simplest is to average the embeddings
 7948 of each word in the sentence, and pass this average through a feedforward neural net-
 7949 work (Iyyer et al., 2015). Convolutional and recurrent neural networks go further, with
 7950 the ability to effectively capturing multiword phenomena such as negation (Kalchbrenner
 7951 et al., 2014; Kim, 2014; Li et al., 2015; Tang et al., 2015). Another approach is to incorpo-
 7952 rate the syntactic structure of the sentence into a **recursive neural network**, in which the
 7953 representation for each syntactic constituent is computed from the representations of its
 7954 children (Socher et al., 2012). However, in many cases, recurrent neural networks perform
 7955 as well or better than recursive networks (Li et al., 2015).

7956 Whether convolutional, recurrent, or recursive, a key question is whether supervised
 7957 sentence representations are task-specific, or whether a single supervised sentence repre-
 7958 sentation model can yield useful performance on other tasks. Wieting et al. (2016b) train a
 7959 variety of sentence embedding models for the task of labeling pairs of sentences as **para-**
 7960 **phrases**. They show that the resulting sentence embeddings give good performance for
 7961 sentiment analysis. The **Stanford Natural Language Inference corpus** classifies sentence
 7962 pairs as **entailments** (the truth of sentence i implies the truth of sentence j), **contradictions**
 7963 (the truth of sentence i implies the falsity of sentence j), and neutral (i neither entails nor
 7964 contradicts j). Sentence embeddings trained on this dataset transfer to a wide range of
 7965 classification tasks (Conneau et al., 2017).

7966 14.8.4 Hybrid distributed-symbolic representations

7967 The power of distributed representations is in their generality: the distributed represen-
 7968 tation of a unit of text can serve as a summary of its meaning, and therefore as the input
 7969 for downstream tasks such as classification, matching, and retrieval. For example, dis-
 7970 tributed sentence representations can be used to recognize the paraphrase relationship
 7971 between closely related sentences like the following:

- 7972 (14.5) a. Donald thanked Vlad profusely.
7973 b. Donald conveyed to Vlad his profound appreciation.
7974 c. Vlad was showered with gratitude by Donald.

7975 Symbolic representations are relatively brittle to this sort of variation, but are better
7976 suited to describe individual entities, the things that they do, and the things that are done
7977 to them. In examples (14.5a)-(14.5c), we not only know that somebody thanked someone
7978 else, but we can make a range of inferences about what has happened between the entities
7979 named *Donald* and *Vlad*. Because distributed representations do not treat entities symbol-
7980 ically, they lack the ability to reason about the roles played by entities across a sentence or
7981 larger discourse.⁵ A hybrid between distributed and symbolic representations might give
7982 the best of both worlds: robustness to the many different ways of describing the same
7983 event, plus the expressiveness to support inferences about entities and the roles that they
7984 play.

7985 A “top-down” hybrid approach is to begin with logical semantics (of the sort de-
7986 scribed in the previous two chapters), and but replace the predefined lexicon with a set
7987 of distributional word clusters (Poon and Domingos, 2009; Lewis and Steedman, 2013). A
7988 “bottom-up” approach is to add minimal symbolic structure to existing distributed repre-
7989 sentations, such as vector representations for each entity (Ji and Eisenstein, 2015; Wiseman
7990 et al., 2016). This has been shown to improve performance on two problems that we will
7991 encounter in the following chapters: classification of **discourse relations** between adjac-
7992 ent sentences (chapter 16; Ji and Eisenstein, 2015), and **coreference resolution** of entity
7993 mentions (chapter 15; Wiseman et al., 2016; Ji et al., 2017). Research on hybrid seman-
7994 tic representations is still in an early stage, and future representations may deviate more
7995 boldly from existing symbolic and distributional approaches.

7996 Additional resources

7997 Turney and Pantel (2010) survey a number of facets of vector word representations, fo-
7998 cusing on matrix factorization methods. Schnabel et al. (2015) highlight problems with
7999 similarity-based evaluations of word embeddings, and present a novel evaluation that
8000 controls for word frequency. Baroni et al. (2014) address linguistic issues that arise in
8001 attempts to combine distributed and compositional representations.

8002 In bilingual and multilingual distributed representations, embeddings are estimated
8003 for translation pairs or tuples, such as (*dog*, *perro*, *chien*). These embeddings can improve
8004 machine translation (Zou et al., 2013; Klementiev et al., 2012), transfer natural language

⁵At a 2014 workshop on semantic parsing, this critique of distributed representations was expressed by Ray Mooney — a leading researcher in computational semantics — in a now well-known quote, “you can’t cram the meaning of a whole sentence into a single vector!”

8005 processing models across languages (Täckström et al., 2012), and make monolingual word
 8006 embeddings more accurate (Faruqui and Dyer, 2014). A typical approach is to learn a pro-
 8007 jection that maximizes the correlation of the distributed representations of each element
 8008 in a translation pair, which can be obtained from a bilingual dictionary. Distributed rep-
 8009 resentations can also be linked to perceptual information, such as image features. Bruni
 8010 et al. (2014) use textual descriptions of images to obtain visual contextual information for
 8011 various words, which supplements traditional distributional context. Image features can
 8012 also be inserted as contextual information in log bilinear language models (Kiros et al.,
 8013 2014), making it possible to automatically generate text descriptions of images.

8014 Exercises

- 8015 1. Prove that the sum of probabilities of paths through a hierarchical softmax tree is
 8016 equal to one.
2. In skipgram word embeddings, the negative sampling objective can be written as,

$$\mathcal{L} = \sum_{i \in \mathcal{V}} \sum_{j \in \mathcal{C}} \text{count}(i, j) \psi(i, j), \quad [14.29]$$

8017 with $\psi(i, j)$ is defined in Equation 14.23.

8018 Suppose we draw the negative samples from the empirical unigram distribution
 8019 $\hat{p}(i) = p_{\text{unigram}}(i)$. First, compute the expectation of \mathcal{L} with respect the negative
 8020 samples, using this probability.

8021 Next, take the derivative of this expectation with respect to the score of a single word
 8022 context pair $\sigma(\mathbf{u}_i \cdot \mathbf{v}_j)$, and solve for the pointwise mutual information $\text{PMI}(i, j)$. You
 8023 should be able to show that at the optimum, the PMI is a simple function of $\sigma(\mathbf{u}_i \cdot \mathbf{v}_j)$
 8024 and the number of negative samples.

8025 (This exercise is part of a proof that shows that skipgram with negative sampling is
 8026 closely related to PMI-weighted matrix factorization.)

- 8027 3. * In Brown clustering, prove that the cluster merge that maximizes the average mu-
 8028 tual information (Equation 14.13) also maximizes the log-likelihood objective (Equa-
 8029 tion 14.12).
4. A simple way to compute a distributed phrase representation is to add up the dis-
 tributed representations of the words in the phrase. Consider a sentiment analysis
 model in which the predicted sentiment is, $\psi(\mathbf{w}) = \theta \cdot (\sum_{m=1}^M \mathbf{x}_m)$, where \mathbf{x}_m is
 the vector representation of word m . Prove that in such a model, the following two

inequalities cannot both hold:

$$\psi(\text{good}) > \psi(\text{not good}) \quad [14.30]$$

$$\psi(\text{bad}) < \psi(\text{not bad}). \quad [14.31]$$

8030 Then construct a similar example pair for the case in which phrase representations
 8031 are the *average* of the word representations.

5. Now let's consider a slight modification to the prediction model in the previous problem:

$$\psi(\mathbf{w}) = \boldsymbol{\theta} \cdot \text{ReLU}\left(\sum_{m=1}^M \mathbf{x}_m\right) \quad [14.32]$$

8032 Show that in this case, it *is* possible to achieve the inequalities above. Your solution
 8033 should provide the weights $\boldsymbol{\theta}$ and the embeddings \mathbf{x}_{good} , \mathbf{x}_{bad} , and \mathbf{x}_{not} .

8034 For the next two problems, download a set of pre-trained word embeddings, such as the
 8035 WORD2VEC or polyglot embeddings.

- 8036 6. Use cosine similarity to find the most similar words to: *dog*, *whale*, *before*, *however*,
 8037 *fabricate*.
- 8038 7. Use vector addition and subtraction to compute target vectors for the analogies below. After computing each target vector, find the top three candidates by cosine
 8039 similarity.
- 8040 • *dog:puppy :: cat:?*
 8041 • *speak:speaker :: sing:?*
 8042 • *France:French :: England:?*
 8043 • *France:wine :: England:?*

8045 The remaining problems will require you to build a classifier and test its properties. Pick a
 8046 text classification dataset, such as the Cornell Movie Review data.⁶ Divide your data into
 8047 training (60%), development (20%), and test sets (20%), if no such division already exists.

- 8048 8. Train a convolutional neural network, with inputs set to pre-trained word embed-
 8049 dings from the previous two problems. Use an additional, fine-tuned embedding
 8050 for out-of-vocabulary words. Train until performance on the development set does
 8051 not improve. You can also use the development set to tune the model architecture,
 8052 such as the convolution width and depth. Report *F-MEASURE* and accuracy, as well
 8053 as training time.

⁶<http://www.cs.cornell.edu/people/pabo/movie-review-data/>

- 8054 9. Now modify your model from the previous problem to fine-tune the word embed-
8055 dings. Report F -MEASURE, accuracy, and training time.
- 8056 10. Try a simpler approach, in which word embeddings in the document are averaged,
8057 and then this average is passed through a feed-forward neural network. Again, use
8058 the development data to tune the model architecture. How close is the accuracy to
8059 the convolutional networks from the previous problems?

8060 **Chapter 15**

8061 **Reference Resolution**

8062 References are one of the most noticeable forms of linguistic ambiguity, afflicting not just
8063 automated natural language processing systems, but also fluent human readers. Warnings
8064 to avoid “ambiguous pronouns” are ubiquitous in manuals and tutorials on writing
8065 style. But referential ambiguity is not limited to pronouns, as shown in the text in Fig-
8066 ure 15.1. Each of the bracketed substrings refers to an entity that is introduced earlier
8067 in the passage. These references include the pronouns *he* and *his*, but also the shortened
8068 name *Cook*, and **nominals** such as *the firm* and *the firm’s biggest growth market*.

8069 **Reference resolution** subsumes several subtasks. This chapter will focus on **corefer-
8070 ence resolution**, which is the task of grouping spans of text that refer to a single underly-
8071 ing entity, or, in some cases, a single event: for example, the spans *Tim Cook*, *he*, and *Cook*
8072 are all **coreferent**. These individual spans are called **mentions**, because they mention an
8073 entity; the entity is sometimes called the **referent**. Each mention has a set of **antecedents**,
8074 which are preceding mentions that are coreferent; for the first mention of an entity, the an-
8075 tecedent set is empty. The task of **pronominal anaphora resolution** requires identifying
8076 only the antecedents of pronouns. In **entity linking**, references are resolved not to other
8077 spans of text, but to entities in a knowledge base. This task is discussed in chapter 17.

8078 Coreference resolution is a challenging problem for several reasons. Resolving differ-
8079 ent types of **referring expressions** requires different types of reasoning: the features and
8080 methods that are useful for resolving pronouns are different from those that are useful
8081 to resolve names and nominals. Coreference resolution involves not only linguistic rea-
8082 soning, but also world knowledge and pragmatics: you may not have known that China
8083 was Apple’s biggest growth market, but it is likely that you effortlessly resolved this ref-
erence while reading the passage in Figure 15.1.¹ A further challenge is that coreference

¹This interpretation is based in part on the assumption that a **cooperative** author would not use the expression *the firm’s biggest growth market* to refer to an entity not yet mentioned in the article (Grice, 1975). **Pragmatics** is the discipline of linguistics concerned with the formalization of such assumptions (Huang,

- (15.1) *[[Apple Inc] Chief Executive Tim Cook] has jetted into [China] for talks with government officials as [he] seeks to clear up a pile of problems in [[the firm] 's biggest growth market] ... [Cook] is on [his] first trip to [the country] since taking over...*
-

Figure 15.1: Running example (Yee and Jones, 2012). Coreferring entity mentions are in brackets.

8085 resolution decisions are often entangled: each mention adds information about the entity,
 8086 which affects other coreference decisions. This means that coreference resolution must
 8087 be addressed as a structure prediction problem. But as we will see, there is no dynamic
 8088 program that allows the space of coreference decisions to be searched efficiently.

8089 **15.1 Forms of referring expressions**

8090 There are three main forms of referring expressions — pronouns, names, and nominals.

8091 **15.1.1 Pronouns**

8092 Pronouns are a closed class of words that are used for references. A natural way to think
 8093 about pronoun resolution is SMASH (Kehler, 2007):

- 8094 • Search for candidate antecedents;
 8095 • Match against hard agreement constraints;
 8096 • And Select using Heuristics, which are “soft” constraints such as recency, syntactic
 8097 prominence, and parallelism.

8098 **Search**

8099 In the search step, candidate antecedents are identified from the preceding text or speech.²
 8100 Any noun phrase can be a candidate antecedent, and pronoun resolution usually requires

2015).

²Pronouns whose referents come later are known as **cataphora**, as in the opening line from a novel by Márquez (1970):

- (15.1) Many years later, as [he] faced the firing squad, [Colonel Aureliano Buendía] was to remember that distant afternoon when [his] father took him to discover ice.

8101 parsing the text to identify all such noun phrases.³ Filtering heuristics can help to prune
 8102 the search space to noun phrases that are likely to be coreferent (Lee et al., 2013; Durrett
 8103 and Klein, 2013). In nested noun phrases, mentions are generally considered to be the
 8104 largest unit with a given **head word** (see § 10.5.2): thus, *Apple Inc. Chief Executive Tim Cook*
 8105 would be included as a mention, but *Tim Cook* would not, since they share the same head
 8106 word, *Cook*.

8107 **Matching constraints for pronouns**

8108 References and their antecedents must agree on semantic features such as number, person,
 8109 gender, and animacy. Consider the pronoun *he* in this passage from the running example:

8110 (15.2) Tim Cook has jetted in for talks with officials as [he] seeks to clear up a pile of
 8111 problems...

8112 The pronoun and possible antecedents have the following features:

- 8113 • *he*: singular, masculine, animate, third person
- 8114 • *officials*: plural, animate, third person
- 8115 • *talks*: plural, inanimate, third person
- 8116 • *Tim Cook*: singular, masculine, animate, third person

8117 The SMASH method searches backwards from *he*, discarding *officials* and *talks* because they
 8118 do not satisfy the agreements constraints.

8119 Another source of constraints comes from syntax — specifically, from the phrase struc-
 8120 ture trees discussed in chapter 10. Consider a parse tree in which both *x* and *y* are phrasal
 8121 constituents. The constituent *x* **c-commands** the constituent *y* iff the first branching node
 8122 above *x* also dominates *y*. For example, in Figure 15.2a, *Abigail* c-commands *her*, because
 8123 the first branching node above *Abigail*, *S*, also dominates *her*. Now, if *x* c-commands *y*,
 8124 **government and binding theory** (Chomsky, 1982) states that *y* can refer to *x* only if it is
 8125 a **reflexive pronoun** (e.g., *herself*). Furthermore, if *y* is a reflexive pronoun, then its an-
 8126 tecedent must c-command it. Thus, in Figure 15.2a, *her* cannot refer to *Abigail*; conversely,
 8127 if we replace *her* with *herself*, then the reflexive pronoun *must* refer to *Abigail*, since this is
 8128 the only candidate antecedent that c-commands it.

8129 Now consider the example shown in Figure 15.2b. Here, *Abigail* does not c-command
 8130 *her*, but *Abigail's mom* does. Thus, *her* can refer to *Abigail* — and we cannot use reflexive

³In the OntoNotes coreference annotations, verbs can also be antecedents, if they are later referenced by nominals (Pradhan et al., 2011):

(15.1) Sales of passenger cars [grew] 22%. [The strong growth] followed year-to-year increases.

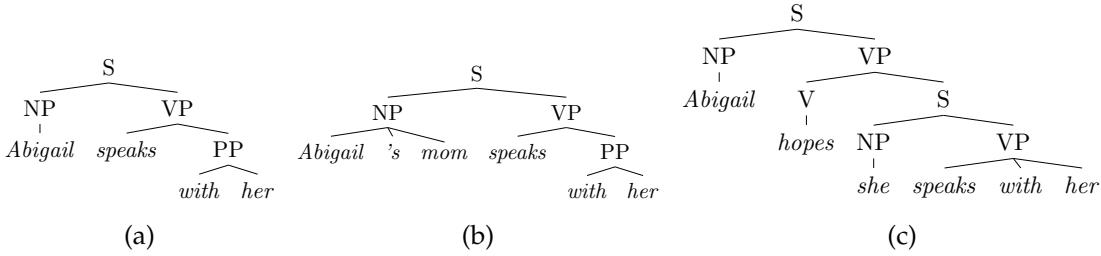


Figure 15.2: In (a), *Abigail* c-commands *her*; in (b), *Abigail* does not c-command *her*, but *Abigail's mom* does; in (c), the scope of *Abigail* is limited by the S non-terminal, so that *she* or *her* can bind to *Abigail*, but not both.

8131 *herself* in this context, unless we are talking about *Abigail*'s mom. However, *her* does not
 8132 have to refer to *Abigail*. Finally, Figure 15.2c shows how these constraints are limited.
 8133 In this case, the pronoun *she* can refer to *Abigail*, because the S non-terminal puts *Abigail*
 8134 outside the domain of *she*. Similarly, *her* can also refer to *Abigail*. But *she* and *her* cannot be
 8135 coreferent, because *she* c-commands *her*.

8136 Heuristics

8137 After applying constraints, heuristics are applied to select among the remaining candidates.
 8138 Recency is a particularly strong heuristic. All things equal, readers will prefer
 8139 the more recent referent for a given pronoun, particularly when comparing referents that
 8140 occur in different sentences. Jurafsky and Martin (2009) offer the following example:

- 8141 (15.3) The doctor found an old map in the captain's chest. Jim found an even older map
 8142 hidden on the shelf. [It] described an island.

8143 Readers are expected to prefer the older map as the referent for the pronoun *it*.

8144 However, subjects are often preferred over objects, and this can contradict the preference
 8145 for recency when two candidate referents are in the same sentence. For example,

- 8146 (15.4) Abigail loaned Lucia a book on Spanish. [She] is always trying to help people.

8147 Here, we may prefer to link *she* to *Abigail* rather than *Lucia*, because of *Abigail*'s position in
 8148 the subject role of the preceding sentence. (Arguably, this preference would not be strong
 8149 enough to select *Abigail* if the second sentence were *She is visiting Valencia next month*.)

8150 A third heuristic is parallelism:

- 8151 (15.5) Abigail loaned Lucia a book on Spanish. Özlem loaned [her] a book on Por-
 8152 tuguese.

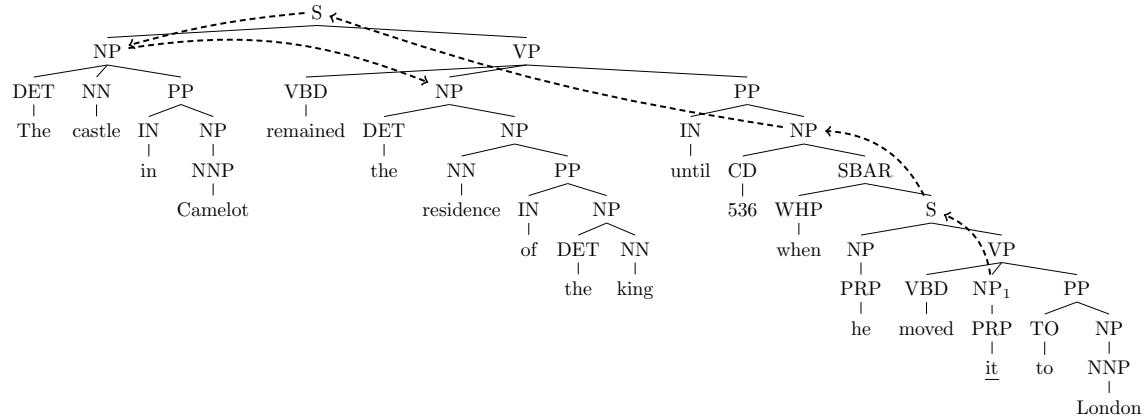


Figure 15.3: Left-to-right breadth-first tree traversal (Hobbs, 1978), indicating that the search for an antecedent for *it* (NP_1) would proceed in the following order: 536; *the castle in Camelot*; *the residence of the king*; *Camelot*; *the king*. Hobbs (1978) proposes semantic constraints to eliminate 536 and *the castle in Camelot* as candidates, since they are unlikely to be the direct object of the verb *move*.

8153 Here *Lucia* is preferred as the referent for *her*, contradicting the preference for the subject
 8154 *Abigail* in the preceding example.

8155 The recency and subject role heuristics can be unified by traversing the document in
 8156 a syntax-driven fashion (Hobbs, 1978): each preceding sentence is traversed breadth-first,
 8157 left-to-right (Figure 15.3). This heuristic successfully handles (15.4): *Abigail* is preferred as
 8158 the referent for *she* because the subject NP is visited first. It also handles (15.3): the older
 8159 map is preferred as the referent for *it* because the more recent sentence is visited first. (An
 8160 alternative unification of recency and syntax is proposed by **centering theory** (Grosz et al.,
 8161 1995), which is discussed in detail in chapter 16.)

8162 In early work on reference resolution, the number of heuristics was small enough that
 8163 a set of numerical weights could be set by hand (Lappin and Leass, 1994). More recent
 8164 work uses machine learning to quantify the importance of each of these factors. However,
 8165 pronoun resolution cannot be completely solved by constraints and heuristics alone. This
 8166 is shown by the classic example pair (Winograd, 1972):

8167 (15.6) The [city council] denied [the protesters] a permit because [they] advocated / feared
 8168 violence.

8169 Without reasoning about the motivations of the city council and protesters, it is unlikely
 8170 that any system could correctly resolve both versions of this example.

8171 **Non-referential pronouns**

8172 While pronouns are generally used for reference, they need not refer to entities. The fol-
 8173 lowing examples show how pronouns can refer to propositions, events, and speech acts.

- 8174 (15.7) a. They told me that I was too ugly for show business, but I didn't believe [it].
 8175 b. Elifsu saw Berthold get angry, and I saw [it] too.
 8176 c. Emmanuel said he worked in security. I suppose [that]'s one way to put it.

8177 These forms of reference are generally not annotated in large-scale coreference resolution
 8178 datasets such as OntoNotes (Pradhan et al., 2011).

8179 Pronouns may also have **generic referents**:

- 8180 (15.8) a. A poor carpenter blames [her] tools.
 8181 b. On the moon, [you] have to carry [your] own oxygen.
 8182 c. Every farmer who owns a donkey beats [it]. (Geach, 1962)

8183 In the OntoNotes dataset, coreference is not annotated for generic referents, even in cases
 8184 like these examples, in which the same generic entity is mentioned multiple times.

8185 Some pronouns do not refer to anything at all:

- 8186 (15.9) a. *[It]'s raining.*
 [Il] pleut. (Fr)
 8187 b. [It] 's money that she's really after.
 8188 c. [It] is too bad that we have to work so hard.

8189 How can we automatically distinguish these usages of *it* from referential pronouns?
 8190 Consider the the difference between the following two examples (Bergsma et al., 2008):

- 8191 (15.10) a. You can make [it] in advance.
 8192 b. You can make [it] in showbiz.

8193 In the second example, the pronoun *it* is non-referential. One way to see this is by substi-
 8194 tuting another pronoun, like *them*, into these examples:

- 8195 (15.11) a. You can make [them] in advance.
 8196 b. ? You can make [them] in showbiz.

8197 The questionable grammaticality of the second example suggests that *it* is not referential.
 8198 Bergsma et al. (2008) operationalize this idea by comparing distributional statistics for the
 8199 *n*-grams around the word *it*, testing how often other pronouns or nouns appear in the
 8200 same context. In cases where nouns and other pronouns are infrequent, the *it* is unlikely
 8201 to be referential.

8202 15.1.2 Proper Nouns

8203 If a proper noun is used as a referring expression, it often corefers with another proper
8204 noun, so that the coreference problem is simply to determine whether the two names
8205 match. Subsequent proper noun references often use a shortened form, as in the running
8206 example (Figure 15.1):

- 8207 (15.12) Apple Inc Chief Executive [Tim Cook] has jetted into China ... [Cook] is on his
8208 first business trip to the country ...

8209 A typical solution for proper noun coreference is to match the syntactic head words
8210 of the reference with the referent. In § 10.5.2, we saw that the head word of a phrase can
8211 be identified by applying head percolation rules to the phrasal parse tree; alternatively,
8212 the head can be identified as the root of the dependency subtree covering the name. For
8213 sequences of proper nouns, the head word will be the final token.

8214 There are a number of caveats to the practice of matching head words of proper nouns.

- 8215 • In the European tradition, family names tend to be more specific than given names,
8216 and family names usually come last. However, other traditions have other practices:
8217 for example, in Chinese names, the family name typically comes first; in Japanese,
8218 honorifics come after the name, as in *Nobu-San* (*Mr. Nobu*).
- 8219 • In organization names, the head word is often not the most informative, as in *Georgia*
8220 *Tech* and *Virginia Tech*. Similarly, *Lebanon* does not refer to the same entity as *Southern Lebanon*, necessitating special rules for the specific case of geographical modi-
8221 fiers (Lee et al., 2011).
- 8223 • Proper nouns can be nested, as in [*the CEO of [Microsoft]*], resulting in head word
8224 match without coreference.

8225 Despite these difficulties, proper nouns are the easiest category of references to re-
8226 solve (Stoyanov et al., 2009). In machine learning systems, one solution is to include a
8227 range of matching features, including exact match, head match, and string inclusion. In
8228 addition to matching features, competitive systems (e.g., Bengtson and Roth, 2008) in-
8229 clude large lists, or **gazetteers**, of acronyms (e.g., *the National Basketball Association/NBA*),
8230 demonym (e.g., *the Israelis/Israel*), and other aliases (e.g., *the Georgia Institute of Technol-*
8231 *ogy/Georgia Tech*).

8232 15.1.3 Nominals

8233 In coreference resolution, noun phrases that are neither pronouns nor proper nouns are
8234 referred to as **nominals**. In the running example (Figure 15.1), nominal references include:
8235 *the firm* (*Apple Inc*); *the firm's biggest growth market* (*China*); and *the country* (*China*).

8236 Nominals are especially difficult to resolve (Denis and Baldridge, 2007; Durrett and
 8237 Klein, 2013), and the examples above suggest why this may be the case: world knowledge
 8238 is required to identify *Apple Inc* as a *firm*, and *China* as a *growth market*. Other difficult
 8239 examples include the use of colloquial expressions, such as coreference between *Clinton*
 8240 *campaign officials* and *the Clinton camp* (Soon et al., 2001).

8241 15.2 Algorithms for coreference resolution

The ground truth training data for coreference resolution is a set of mention sets, where all mentions within each set refer to a single entity.⁴ In the running example from Figure 15.1, the ground truth coreference annotation is:

$$c_1 = \{Apple\ Inc_{1:2}, the\ firm_{27:28}\} \quad [15.1]$$

$$c_2 = \{Apple\ Inc\ Chief\ Executive\ Tim\ Cook_{1:6}, he_{17}, Cook_{33}, his_{36}\} \quad [15.2]$$

$$c_3 = \{China_{10}, the\ firm\ 's\ biggest\ growth\ market_{27:32}, the\ country_{40:41}\} \quad [15.3]$$

8242 Each row specifies the token spans that mention an entity. (“Singleton” entities, which are
 8243 mentioned only once (e.g., *talks*, *government officials*), are excluded from the annotations.)
 8244 Equivalently, if given a set of M mentions, $\{m_i\}_{i=1}^M$, each mention i can be assigned to a
 8245 cluster z_i , where $z_i = z_j$ if i and j are coreferent. The cluster assignments z are invariant
 8246 under permutation. The unique clustering associated with the assignment z is written
 8247 $c(z)$.

8248 Coreference resolution can thus be viewed as a structure prediction problem, involving
 8249 two subtasks: identifying which spans of text mention entities, and then clustering
 8250 those spans.

8251 **Mention identification** The task of identifying mention spans for coreference resolution
 8252 is often performed by applying a set of heuristics to the phrase structure parse of each
 8253 sentence. A typical approach is to start with all noun phrases and named entities, and
 8254 then apply filtering rules to remove nested noun phrases with the same head (e.g., [*Apple*
 8255 CEO [*Tim Cook*]]), numeric entities (e.g., [*100 miles*], [*97%*]), non-referential *it*, etc (Lee
 8256 et al., 2013; Durrett and Klein, 2013). In general, these deterministic approaches err in
 8257 favor of recall, since the mention clustering component can choose to ignore false positive
 8258 mentions, but cannot recover from false negatives. An alternative is to consider all spans
 8259 (up to some finite length) as candidate mentions, performing mention identification and
 8260 clustering jointly (Daumé III and Marcu, 2005; Lee et al., 2017).

⁴In many annotations, the term **markable** is used to refer to spans of text that can *potentially* mention an entity. The set of markables includes non-referential pronouns, which does not mention any entity. Part of the job of the coreference system is to avoid incorrectly linking these non-referential markables to any mention chains.

8261 **Mention clustering** The subtask of mention clustering will be the focus of the remainder
 8262 of this chapter. There are two main classes of models. In *mention-based models*, the scoring
 8263 function for a coreference clustering decomposes over pairs of mentions. These pairwise
 8264 decisions are then aggregated, using a clustering heuristic. Mention-based coreference
 8265 clustering can be treated as a fairly direct application of supervised classification or rank-
 8266 ing. However, the mention-pair locality assumption can result in incoherent clusters, like
 8267 $\{\text{Hillary Clinton} \leftarrow \text{Clinton} \leftarrow \text{Mr Clinton}\}$, in which the pairwise links score well, but the
 8268 overall result is unsatisfactory. *Entity-based models* address this issue by scoring entities
 8269 holistically. This can make inference more difficult, since the number of possible entity
 8270 groupings is exponential in the number of mentions.

8271 15.2.1 Mention-pair models

8272 In the **mention-pair model**, a binary label $y_{i,j} \in \{0, 1\}$ is assigned to each pair of mentions
 8273 (i, j), where $i < j$. If i and j corefer ($z_i = z_j$), then $y_{i,j} = 1$; otherwise, $y_{i,j} = 0$. The
 8274 mention *he* in Figure 15.1 is preceded by five other mentions: (1) *Apple Inc*; (2) *Apple Inc*
 8275 *Chief Executive Tim Cook*; (3) *China*; (4) *talks*; (5) *government officials*. The correct mention
 8276 pair labeling is $y_{2,6} = 1$ and $y_{i \neq 2,6} = 0$ for all other i . If a mention j introduces a new entity,
 8277 such as mention 3 in the example, then $y_{i,j} = 0$ for all i . The same is true for “mentions”
 8278 that do not refer to any entity, such as non-referential pronouns. If mention j refers to an
 8279 entity that has been mentioned more than once, then $y_{i,j} = 1$ for all $i < j$ that mention the
 8280 referent.

8281 By transforming coreference into a set of binary labeling problems, the mention-pair
 8282 model makes it possible to apply an off-the-shelf binary classifier (Soon et al., 2001). This
 8283 classifier is applied to each mention j independently, searching backwards from j until
 8284 finding an antecedent i which corefers with j with high confidence. After identifying a
 8285 single **antecedent**, the remaining mention pair labels can be computed by transitivity: if
 8286 $y_{i,j} = 1$ and $y_{j,k} = 1$, then $y_{i,k} = 1$.

8287 Since the ground truth annotations give entity chains c but not individual mention-
 8288 pair labels y , an additional heuristic must be employed to convert the labeled data into
 8289 training examples for classification. A typical approach is to generate at most one pos-
 8290 itive labeled instance $y_{a_j,j} = 1$ for mention j , where a_j is the index of the most recent
 8291 antecedent, $a_j = \max\{i : i < j \wedge z_i = z_j\}$. Negative labeled instances are generated for
 8292 all for all $i \in \{a_j + 1, \dots, j\}$. In the running example, the most recent antecedent of the
 8293 pronoun *he* is $a_6 = 2$, so the training data would be $y_{2,6} = 1$ and $y_{3,6} = y_{4,6} = y_{5,6} = 0$.
 8294 The variable $y_{1,6}$ is not part of the training data, because the first mention appears before
 8295 the true antecedent $a_6 = 2$.

8296 **15.2.2 Mention-ranking models**

In **mention ranking** (Denis and Baldridge, 2007), the classifier learns to identify a single antecedent $a_i \in \{\epsilon, 1, 2, \dots, i-1\}$ for each referring expression i ,

$$\hat{a}_i = \operatorname{argmax}_{a \in \{\epsilon, 1, 2, \dots, i-1\}} \psi_M(a, i), \quad [15.4]$$

8297 where $\psi_M(a, i)$ is a score for the mention pair (a, i) . If $a = \epsilon$, then mention i does not refer
 8298 to any previously-introduced entity — it is not **anaphoric**. Mention-ranking is similar to
 8299 the mention-pair model, but all candidates are considered simultaneously, and at most
 8300 a single antecedent is selected. The mention-ranking model explicitly accounts for the
 8301 possibility that mention i is not anaphoric, through the score $\psi_M(\epsilon, i)$. The determination
 8302 of anaphoricity can be made by a special classifier in a preprocessing step, so that non- ϵ
 8303 antecedents are identified only for spans that are determined to be anaphoric (Denis and
 8304 Baldridge, 2008).

8305 As a learning problem, ranking can be trained using the same objectives as in dis-
 8306 criminative classification. For each mention i , we can define a gold antecedent a_i^* , and an
 8307 associated loss, such as the hinge loss, $\ell_i = (1 - \psi_M(a_i^*, i) + \psi_M(\hat{a}, i))_+$ or the negative
 8308 log-likelihood, $\ell_i = -\log p(a_i^* | i; \theta)$. (For more on learning to rank, see § 17.1.1.) But as
 8309 with the mention-pair model, there is a mismatch between the labeled data, which comes
 8310 in the form of mention sets, and the desired supervision, which would indicate the spe-
 8311 cific antecedent of each mention. The antecedent variables $\{a_i\}_{i=1}^M$ relate to the mention
 8312 sets in a many-to-one mapping: each set of antecedents induces a single clustering, but a
 8313 clustering can correspond to many different settings of antecedent variables.

A heuristic solution is to set $a_i^* = \max\{j : j < i \wedge z_j = z_i\}$, the most recent mention in
 the same cluster as i . But the most recent mention may not be the most informative: in the
 running example, the most recent antecedent of the mention *Cook* is the pronoun *he*, but
 a more useful antecedent is the earlier mention *Apple Inc Chief Executive Tim Cook*. Rather
 than selecting a specific antecedent to train on, the antecedent can be treated as a latent
 variable, in the manner of the **latent variable perceptron** from § 12.4.2 (Fernandes et al.,
 2014):

$$\hat{\mathbf{a}} = \operatorname{argmax}_{\mathbf{a}} \sum_{i=1}^M \psi_M(a_i, i) \quad [15.5]$$

$$\mathbf{a}^* = \operatorname{argmax}_{\mathbf{a} \in \mathcal{A}(c)} \sum_{i=1}^M \psi_M(a_i, i) \quad [15.6]$$

$$\theta \leftarrow \theta + \sum_{i=1}^M \frac{\partial L}{\partial \theta} \psi_M(a_i^*, i) - \sum_{i=1}^M \frac{\partial L}{\partial \theta} \psi_M(\hat{a}_i, i) \quad [15.7]$$

where $\mathcal{A}(c)$ is the set of antecedent structures that is compatible with the ground truth coreference clustering c . Another alternative is to sum over all the conditional probabilities of antecedent structures that are compatible with the ground truth clustering (Durrett and Klein, 2013; Lee et al., 2017). For the set of mention m , we compute the following probabilities:

$$p(c | m) = \sum_{a \in \mathcal{A}(c)} p(a | m) = \sum_{a \in \mathcal{A}(c)} \prod_{i=1}^M p(a_i | i, m) \quad [15.8]$$

$$p(a_i | i, m) = \frac{\exp(\psi_M(a_i, i))}{\sum_{a' \in \{\epsilon, 1, 2, \dots, i-1\}} \exp(\psi_M(a', i))}. \quad [15.9]$$

- 8314 This objective rewards models that assign high scores for all valid antecedent structures.
 8315 In the running example, this would correspond to summing the probabilities of the two
 8316 valid antecedents for *Cook, he* and *Apple Inc Chief Executive Tim Cook*. In one of the exer-
 8317 cises, you will compute the number of valid antecedent structures for a given clustering.

8318 15.2.3 Transitive closure in mention-based models

A problem for mention-based models is that individual mention-level decisions may be incoherent. Consider the following mentions:

$$m_1 = \text{Hillary Clinton} \quad [15.10]$$

$$m_2 = \text{Clinton} \quad [15.11]$$

$$m_3 = \text{Bill Clinton} \quad [15.12]$$

- 8319 A mention-pair system might predict $\hat{y}_{1,2} = 1, \hat{y}_{2,3} = 1, \hat{y}_{1,3} = 0$. Similarly, a mention-
 8320 ranking system might choose $\hat{a}_2 = 1$ and $\hat{a}_3 = 2$. Logically, if mentions 1 and 3 are both
 8321 coreferent with mention 2, then all three mentions must refer to the same entity. This
 8322 constraint is known as **transitive closure**.

- 8323 Transitive closure can be applied *post hoc*, revising the independent mention-pair or
 8324 mention-ranking decisions. However, there are many possible ways to enforce transitive
 8325 closure: in the example above, we could set $\hat{y}_{1,3} = 1$, or $\hat{y}_{1,2} = 0$, or $\hat{y}_{2,3} = 0$. For docu-
 8326 ments with many mentions, there may be many violations of transitive closure, and many
 8327 possible fixes. Transitive closure can be enforced by always adding edges, so that $\hat{y}_{1,3} = 1$
 8328 is preferred (e.g., Soon et al., 2001), but this can result in overclustering, with too many
 8329 mentions grouped into too few entities.

Mention-pair coreference resolution can be viewed as a constrained optimization prob-

lem,

$$\max_{\mathbf{y} \in \{0,1\}^M} \sum_{j=1}^M \sum_{i=1}^j \psi_M(i, j) \times y_{i,j}$$

s.t. $y_{i,j} + y_{j,k} - 1 \leq y_{i,k}, \quad \forall i < j < k,$

with the constraint enforcing transitive closure. This constrained optimization problem is equivalent to graph partitioning with positive and negative edge weights: construct a graph where the nodes are mentions, and the edges are the pairwise scores $\psi_M(i, j)$; the goal is to partition the graph so as to maximize the sum of the edge weights between all nodes within the same partition (McCallum and Wellner, 2004). This problem is NP-hard, motivating approximations such as correlation clustering (Bansal et al., 2004) and **integer linear programming** (Klenner, 2007; Finkel and Manning, 2008, also see § 13.2.2).

15.2.4 Entity-based models

A weakness of mention-based models is that they treat coreference resolution as a classification or ranking problem, when it is really a clustering problem: the goal is to group the mentions together into clusters that correspond to the underlying entities. Entity-based approaches attempt to identify these clusters directly. Such methods require a scoring function at the entity level, measuring whether each set of mentions is internally consistent. Coreference resolution can then be viewed as the following optimization,

$$\max_{\mathbf{z}} \sum_{e=1} \psi_E(\{i : z_i = e\}), \quad [15.13]$$

where z_i indicates the entity referenced by mention i , and $\psi_E(\{i : z_i = e\})$ is a scoring function applied to all mentions i that are assigned to entity e .

Entity-based coreference resolution is conceptually similar to the unsupervised clustering problems encountered in chapter 5: the goal is to obtain clusters of mentions that are internally coherent. The number of possible clusterings of n items is the **Bell number**, which is defined by the following recurrence (Bell, 1934; Luo et al., 2004),

$$B_n = \sum_{k=0}^{n-1} B_k \binom{n-1}{k} B_0 = \quad B_1 = 1. \quad [15.14]$$

This recurrence is illustrated by the Bell tree, which is applied to a short coreference problem in Figure 15.4. The Bell number B_n grows exponentially with n , making exhaustive search of the space of clusterings impossible. For this reason, entity-based coreference resolution typically involves incremental search, in which clustering decisions are based on local evidence, in the hope of approximately optimizing the full objective in Equation 15.13. This approach is sometimes called **cluster ranking**, in contrast to mention ranking.

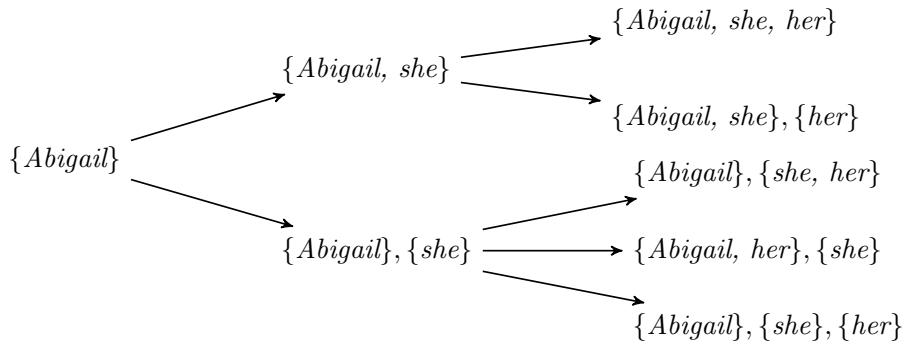


Figure 15.4: The Bell Tree for the sentence *Abigail hopes she speaks with her*. Which paths are excluded by the syntactic constraints mentioned in § 15.1.1?

8347 ***Generative models of coreference** Entity-based coreference can be approached through
 8348 probabilistic **generative models**, in which the mentions in the document are conditioned
 8349 on a set of latent entities (Haghghi and Klein, 2007, 2010). An advantage of these meth-
 8350 ods is that they can be learned from unlabeled data (Poon and Domingos, 2008, e.g.,); a
 8351 disadvantage is that probabilistic inference is required not just for learning, but also for
 8352 prediction. Furthermore, generative models require independence assumptions that are
 8353 difficult to apply in coreference resolution, where the diverse and heterogeneous features
 8354 do not admit an easy decomposition into mutually independent subsets.

8355 Incremental cluster ranking

8356 The SMASH method (§ 15.1.1) can be extended to entity-based coreference resolution by
 8357 building up coreference clusters while moving through the document (Cardie and Wagstaff,
 8358 1999). At each mention, the algorithm iterates backwards through possible antecedent
 8359 clusters; but unlike SMASH, a cluster is selected only if *all* members of its cluster are com-
 8360 patible with the current mention. As mentions are added to a cluster, so are their features
 8361 (e.g., gender, number, animacy). In this way, incoherent chains like *{Hillary Clinton, Clinton, Bill Clinton}*
 8362 can be avoided. However, an incorrect assignment early in the document — a **search error**
 8363 — might lead to a cascade of errors later on.

8364 More sophisticated search strategies can help to ameliorate the risk of search errors.
 8365 One approach is **beam search** (first discussed in § 11.3), in which a set of hypotheses is
 8366 maintained throughout search. Each hypothesis represents a path through the Bell tree
 8367 (Figure 15.4). Hypotheses are “expanded” either by adding the next mention to an ex-
 8368 isting cluster, or by starting a new cluster. Each expansion receives a score, based on
 8369 Equation 15.13, and the top K hypotheses are kept on the beam as the algorithm moves
 8370 to the next step.

8371 Incremental cluster ranking can be made more accurate by performing multiple passes
 8372 over the document, applying rules (or “sieves”) with increasing recall and decreasing
 8373 precision at each pass (Lee et al., 2013). In the early passes, coreference links are pro-
 8374 posed only between mentions that are highly likely to corefer (e.g., exact string match
 8375 for full names and nominals). Information can then be shared among these mentions,
 8376 so that when more permissive matching rules are applied later, agreement is preserved
 8377 across the entire cluster. For example, in the case of *{Hillary Clinton, Clinton, she}*, the
 8378 name-matching sieve would link *Clinton* and *Hillary Clinton*, and the pronoun-matching
 8379 sieve would then link *she* to the combined cluster. A deterministic multi-pass system
 8380 won nearly every track of the 2011 CoNLL shared task on coreference resolution (Prad-
 8381 han et al., 2011). Given the dominance of machine learning in virtually all other areas
 8382 of natural language processing — and more than fifteen years of prior work on machine
 8383 learning for coreference — this was a surprising result, even if learning-based methods
 8384 have subsequently regained the upper hand (e.g., Lee et al., 2018, the state of the art at the
 8385 time of this writing).

8386 **Incremental perceptron**

Incremental coreference resolution can be learned with the **incremental perceptron**, as described in § 11.3.2. At mention i , each hypothesis on the beam corresponds to a clustering of mentions $1 \dots i - 1$, or equivalently, a path through the Bell tree up to position $i - 1$. As soon as none of the hypotheses on the beam are compatible with the gold coreference clustering, a perceptron update is made (Daumé III and Marcu, 2005). For concreteness, consider a linear cluster ranking model,

$$\psi_E(\{i : z_i = e\}) = \sum_{i:z_i=e} \theta \cdot f(i, \{j : j < i \wedge z_j = e\}), \quad [15.15]$$

8387 where the score for each cluster is computed as the sum of scores of all mentions that are
 8388 linked into the cluster, and $f(i, \emptyset)$ is a set of features for the non-anaphoric mention that
 8389 initiates the cluster.

8390 Using Figure 15.4 as an example, suppose that the ground truth is,

$$c^* = \{Abigail, her\}, \{she\}, \quad [15.16]$$

8391 but that with a beam of size one, the learner reaches the hypothesis,

$$\hat{c} = \{Abigail, she\}. \quad [15.17]$$

This hypothesis is incompatible with c^* , so an update is needed:

$$\theta \leftarrow \theta + f(c^*) - f(\hat{c}) \quad [15.18]$$

$$= \theta + (f(Abigail, \emptyset) + f(she, \emptyset)) - (f(Abigail, \emptyset) + f(she, \{Abigail\})) \quad [15.19]$$

$$= \theta + f(she, \emptyset) - f(she, \{Abigail\}). \quad [15.20]$$

8392 This style of incremental update can also be applied to a margin loss between the gold
 8393 clustering and the top clustering on the beam. By backpropagating from this loss, it is also
 8394 possible to train a more complicated scoring function, such as a neural network in which
 8395 the score for each entity is a function of embeddings for the entity mentions (Wiseman
 8396 et al., 2015).

8397 **Reinforcement learning**

8398 **Reinforcement learning** is a topic worthy of a textbook of its own (Sutton and Barto,
 8399 1998),⁵ so this section will provide only a very brief overview, in the context of coreference
 8400 resolution. A stochastic **policy** assigns a probability to each possible **action**, conditional
 8401 on the context. The goal is to learn a policy that achieves a high expected reward, or
 8402 equivalently, a low expected cost.

8403 In incremental cluster ranking, a complete clustering on M mentions can be produced
 8404 by a sequence of M actions, in which the action z_i either merges mention i with an existing
 8405 cluster or begins a new cluster. We can therefore create a stochastic policy using the cluster
 8406 scores (Clark and Manning, 2016),

$$\Pr(z_i = e; \boldsymbol{\theta}) = \frac{\exp \psi_E(i \cup \{j : z_j = e\}; \boldsymbol{\theta})}{\sum_{e'} \exp \psi_E(i \cup \{j : z_j = e'\}; \boldsymbol{\theta})}, \quad [15.21]$$

8407 where $\psi_E(i \cup \{j : z_j = e\}; \boldsymbol{\theta})$ is the score under parameters $\boldsymbol{\theta}$ for assigning mention i to
 8408 cluster e . This score can be an arbitrary function of the mention i , the cluster e and its
 8409 (possibly empty) set of mentions; it can also include the history of actions taken thus far.

8410 If a policy assigns probability $p(c; \boldsymbol{\theta})$ to clustering c , then its expected loss is,

$$L(\boldsymbol{\theta}) = \sum_{c \in \mathcal{C}(\mathbf{m})} p_{\boldsymbol{\theta}}(c) \times \ell(c), \quad [15.22]$$

8411 where $\mathcal{C}(\mathbf{m})$ is the set of possible clusterings for mentions \mathbf{m} . The loss $\ell(c)$ can be based on
 8412 any arbitrary scoring function, including the complex evaluation metrics used in corefer-
 8413 ence resolution (see § 15.4). This is an advantage of reinforcement learning, which can be
 8414 trained directly on the evaluation metric — unlike traditional supervised learning, which
 8415 requires a loss function that is differentiable and decomposable across individual deci-
 8416 sions.

Rather than summing over the exponentially many possible clusterings, we can ap-
 proximate the expectation by sampling trajectories of actions, $\mathbf{z} = (z_1, z_2, \dots, z_M)$, from

⁵A draft of the second edition can be found here: <http://incompleteideas.net/book/the-book-2nd.html>. Reinforcement learning has been used in spoken dialogue systems (Walker, 2000) and text-based game playing (Branavan et al., 2009), and was applied to coreference resolution by Clark and Manning (2015).

the current policy. Each action z_i corresponds to a step in the Bell tree: adding mention m_i to an existing cluster, or forming a new cluster. Each trajectory \mathbf{z} corresponds to a single clustering \mathbf{c} , and so we can write the loss of an action sequence as $\ell(\mathbf{c}(\mathbf{z}))$. The **policy gradient** algorithm computes the gradient of the expected loss as an expectation over trajectories (Sutton et al., 2000),

$$\frac{\partial}{\partial \theta} L(\theta) = E_{\mathbf{z} \sim \mathcal{Z}(\mathbf{m})} \ell(\mathbf{c}(\mathbf{z})) \sum_{i=1}^M \frac{\partial}{\partial \theta} \log p(z_i | \mathbf{z}_{1:i-1}, \mathbf{m}) \quad [15.23]$$

$$\approx \frac{1}{K} \sum_{k=1}^K \ell(\mathbf{c}(\mathbf{z}^{(k)})) \sum_{i=1}^M \frac{\partial}{\partial \theta} \log p(z_i^{(k)} | \mathbf{z}_{1:i-1}^{(k)}, \mathbf{m}), \quad [15.24]$$

where each action sequence $\mathbf{z}^{(k)}$ is sampled from the current policy. Unlike the incremental perceptron, an update is not made until the complete action sequence is available.

8419 Learning to search

8420 Policy gradient can suffer from high variance: while the average loss over K samples is
 8421 asymptotically equal to the expected reward of a given policy, this estimate may not be
 8422 accurate unless K is very large. This can make it difficult to allocate credit and blame to
 8423 individual actions. In **learning to search**, this problem is addressed through the addition
 8424 of an **oracle** policy, which is known to receive zero or small loss. The oracle policy can be
 8425 used in two ways:

- 8426 • The oracle can be used to generate partial hypotheses that are likely to score well,
 8427 by generating i actions from the initial state. These partial hypotheses are then used
 8428 as starting points for the learned policy. This is known as **roll-in**.
- 8429 • The oracle can be used to compute the minimum possible loss from a given state, by
 8430 generating $M - i$ actions from the current state until completion. This is known as
 8431 **roll-out**.

8432 The oracle can be combined with the existing policy during both roll-in and roll-out, sam-
 8433 pling actions from each policy (Daumé III et al., 2009). One approach is to gradually
 8434 decrease the number of actions drawn from the oracle over the course of learning (Ross
 8435 et al., 2011).

8436 In the context of entity-based coreference resolution, Clark and Manning (2016) use
 8437 the learned policy for roll-in and the oracle policy for roll-out. Algorithm 17 shows how
 8438 the gradients on the policy weights are computed in this case. In this application, the
 8439 oracle is “noisy”, because it selects the action that minimizes only the *local* loss — the
 8440 accuracy of the coreference clustering up to mention i — rather than identifying the action
 8441 sequence that will lead to the best final coreference clustering on the entire document.

Algorithm 17 Learning to search for entity-based coreference resolution

```

1: procedure COMPUTE-GRADIENT(mentions  $m$ , loss function  $\ell$ , parameters  $\theta$ )
2:    $L(\theta) \leftarrow 0$ 
3:    $z \sim p(z | m; \theta)$                                  $\triangleright$  Sample a trajectory from the current policy
4:   for  $i \in \{1, 2, \dots, M\}$  do
5:     for action  $z \in \mathcal{Z}(z_{1:i-1}, m)$  do           $\triangleright$  All possible actions after history  $z_{1:i-1}$ 
6:        $h \leftarrow z_{1:i-1} \oplus z$                        $\triangleright$  Concatenate history  $z_{1:i-1}$  with action  $z$ 
7:       for  $j \in \{i+1, i+2, \dots, M\}$  do            $\triangleright$  Roll-out
8:          $h_j \leftarrow \operatorname{argmin}_h \ell(h_{1:j-1} \oplus h)$      $\triangleright$  Oracle selects action with minimum loss
9:        $L(\theta) \leftarrow L(\theta) + p(z | z_{1:i-1}, m; \theta) \times \ell(h)$        $\triangleright$  Update expected loss
10:      return  $\frac{\partial}{\partial \theta} L(\theta)$ 

```

8442 When learning from noisy oracles, it can be helpful to mix in actions from the current
 8443 policy with the oracle during roll-out (Chang et al., 2015).

8444 15.3 Representations for coreference resolution

8445 Historically, coreference resolution has employed an array of hand-engineered features
 8446 to capture the linguistic constraints and preferences described in § 15.1 (Soon et al., 2001).
 8447 Later work has documented the utility of lexical and bilexical features on mention pairs (Björkelund
 8448 and Nugues, 2011; Durrett and Klein, 2013). The most recent and successful methods re-
 8449 place many (but not all) of these features with distributed representations of mentions
 8450 and entities (Wiseman et al., 2015; Clark and Manning, 2016; Lee et al., 2017).

8451 15.3.1 Features

8452 Coreference features generally rely on a preprocessing pipeline to provide part-of-speech
 8453 tags and phrase structure parses. This pipeline makes it possible to design features that
 8454 capture many of the phenomena from § 15.1, and is also necessary for typical approaches
 8455 to mention identification. However, the pipeline may introduce errors that propagate
 8456 to the downstream coreference clustering system. Furthermore, the existence of such
 8457 a pipeline presupposes resources such as treebanks, which do not exist for many lan-
 8458 guages.⁶

⁶The Universal Dependencies project has produced dependency treebanks for more than sixty languages. However, coreference features and mention detection are generally based on phrase structure trees, which exist for roughly two dozen languages. A list is available here: <https://en.wikipedia.org/wiki/Treebank>

8459 **Mention features**

8460 Features of individual mentions can help to predict anaphoricity. In systems where men-
 8461 tion detection is performed jointly with coreference resolution, these features can also
 8462 predict whether a span of text is likely to be a mention. For mention i , typical features
 8463 include:

8464 **Mention type.** Each span can be identified as a pronoun, name, or nominal, using the
 8465 part-of-speech of the head word of the mention: both the Penn Treebank and Uni-
 8466 versal Dependencies tagsets (§ 8.1.1) include tags for pronouns and proper nouns,
 8467 and all other heads can be marked as nominals (Haghghi and Klein, 2009).

8468 **Mention width.** The number of tokens in a mention is a rough predictor of its anaphorici-
 8469 ty, with longer mentions being less likely to refer back to previously-defined enti-
 8470 ties.

8471 **Lexical features.** The first, last, and head words can help to predict anaphoricity; they are
 8472 also useful in conjunction with features such as mention type and part-of-speech,
 8473 providing a rough measure of agreement (Björkelund and Nugues, 2011). The num-
 8474 ber of lexical features can be very large, so it can be helpful to select only frequently-
 8475 occurring features (Durrett and Klein, 2013).

8476 **Morphosyntactic features.** These features include the part-of-speech, number, gender,
 8477 and dependency ancestors.

8478 The features for mention i and candidate antecedent a can be conjoined, producing
 8479 joint features that can help to assess the compatibility of the two mentions. For example,
 8480 Durrett and Klein (2013) conjoin each feature with the mention types of the anaphora
 8481 and the antecedent. Coreference resolution corpora such as ACE and OntoNotes contain
 8482 documents from various genres. By conjoining the genre with other features, it is possible
 8483 to learn genre-specific feature weights.

8484 **Mention-pair features**

8485 For any pair of mentions i and j , typical features include:

8486 **Distance.** The number of intervening tokens, mentions, and sentences between i and j
 8487 can all be used as distance features. These distances can be computed on the surface
 8488 text, or on a transformed representation reflecting the breadth-first tree traversal
 8489 (Figure 15.3). Rather than using the distances directly, they are typically binned,
 8490 creating binary features.

8491 **String match.** A variety of string match features can be employed: exact match, suffix
8492 match, head match, and more complex matching rules that disregard irrelevant
8493 modifiers (Soon et al., 2001).

8494 **Compatibility.** Building on the model, features can measure the anaphor and antecedent
8495 agree with respect to morphosyntactic attributes such as gender, number, and ani-
8496 macy.

8497 **Nesting.** If one mention is nested inside another (e.g., *[The President of [France]]*), they
8498 generally cannot corefer.

8499 **Same speaker.** For documents with quotations, such as news articles, personal pronouns
8500 can be resolved only by determining the speaker for each mention (Lee et al., 2013).
8501 Coreference is also more likely between mentions from the same speaker.

8502 **Gazetteers.** These features indicate that the anaphor and candidate antecedent appear in
8503 a gazetteer of acronyms (e.g., *USA/United States, GATech/Georgia Tech*), demonymns
8504 (e.g., *Israel/Israeli*), or other aliases (e.g., *Knickerbockers/New York Knicks*).

8505 **Lexical semantics.** These features use a lexical resource such as WORDNET to determine
8506 whether the head words of the mentions are related through synonymy, antonymy,
8507 and hypernymy (§ 4.2).

8508 **Dependency paths.** The dependency path between the anaphor and candidate antecedent
8509 can help to determine whether the pair can corefer, under the government and bind-
8510 ing constraints described in § 15.1.1.

8511 Comprehensive lists of mention-pair features are offered by Bengtson and Roth (2008) and
8512 Rahman and Ng (2011). Neural network approaches use far fewer mention-pair features:
8513 for example, Lee et al. (2017) include only speaker, genre, distance, and mention width
8514 features.

8515 **Semantics** In many cases, coreference seems to require knowledge and semantic infer-
8516 ences, as in the running example, where we link *China* with a *country* and a *growth mar-
8517 ket*. Some of this information can be gleaned from WORDNET, which defines a graph
8518 over **synsets** (see § 4.2). For example, one of the synsets of *China* is an instance of an
8519 *Asian_nation#1*, which in turn is a hyponym of *country#2*, a synset that includes
8520 *country*.⁷ Such paths can be used to measure the similarity between concepts (Pedersen
8521 et al., 2004), and this similarity can be incorporated into coreference resolution as a fea-
8522 ture (Ponzetto and Strube, 2006). Similar ideas can be applied to knowledge graphs in-
8523 duced from Wikipedia (Ponzetto and Strube, 2007). But while such approaches improve

⁷teletype font is used to indicate wordnet synsets, and *italics* is used to indicate strings.

8524 relatively simple classification-based systems, they have proven less useful when added
 8525 to the current generation of techniques.⁸ For example, Durrett and Klein (2013) employ
 8526 a range of semantics-based features — WordNet synonymy and hypernymy relations on
 8527 head words, named entity types (e.g., person, organization), and unsupervised cluster-
 8528 ing over nominal heads — but find that these features give minimal improvement over a
 8529 baseline system using surface features.

8530 **Entity features**

8531 Many of the features for entity-mention coreference are generated by aggregating mention-
 8532 pair features over all mentions in the candidate entity (Culotta et al., 2007; Rahman and
 8533 Ng, 2011). Specifically, for each binary mention-pair feature $f(i, j)$, we compute the fol-
 8534 lowing entity-mention features for mention i and entity $e = \{j : j < i \wedge z_j = e\}$.

- 8535 • ALL-TRUE: Feature $f(i, j)$ holds for all mentions $j \in e$.
- 8536 • MOST-TRUE: Feature $f(i, j)$ holds for at least half and fewer than all mentions $j \in e$.
- 8537 • MOST-FALSE: Feature $f(i, j)$ holds for at least one and fewer than half of all men-
 8538 tions $j \in e$.
- 8539 • NONE: Feature $f(i, j)$ does not hold for any mention $j \in e$.

8540 For scalar mention-pair features (e.g., distance features), aggregation can be performed by
 8541 computing the minimum, maximum, and median values across all mentions in the cluster.
 8542 Additional entity-mention features include the number of mentions currently clustered in
 8543 the entity, and ALL-X and MOST-X features for each mention type.

8544 **15.3.2 Distributed representations of mentions and entities**

8545 Recent work has emphasized distributed representations of both mentions and entities.
 8546 One potential advantage is that pre-trained embeddings could help to capture the se-
 8547 mantic compatibility underlying nominal coreference, helping with difficult cases like
 8548 (*Apple, the firm*) and (*China, the firm's biggest growth market*). Furthermore, a distributed
 8549 representation of entities can be trained to capture semantic features that are added by
 8550 each mention.

8551 **Mention embeddings**

8552 Entity mentions can be embedded into a vector space, providing the base layer for neural
 8553 networks that score coreference decisions (Wiseman et al., 2015).

⁸This point was made by Michael Strube at a 2015 workshop, noting that as the quality of the machine learning models in coreference has improved, the benefit of including semantics has become negligible.

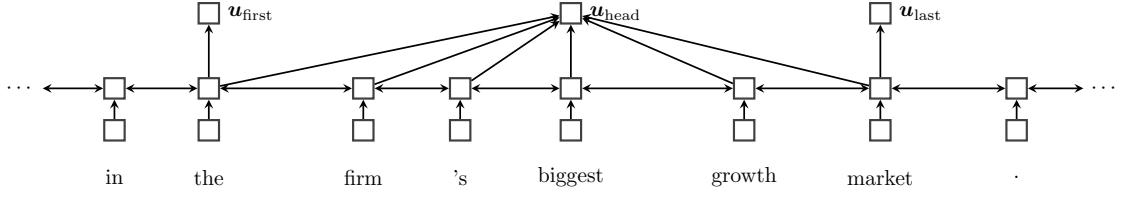


Figure 15.5: A bidirectional recurrent model of mention embeddings. The mention is represented by its first word, its last word, and an estimate of its head word, which is computed from a weighted average (Lee et al., 2017).

8554 **Constructing the mention embedding** Various approaches for embedding multiword
 8555 units can be applied (see § 14.8). Figure 15.5 shows a recurrent neural network approach,
 8556 which begins by running a bidirectional LSTM over the entire text, obtaining hidden states
 8557 from the left-to-right and right-to-left passes, $\mathbf{h}_m = [\overleftarrow{\mathbf{h}}_m; \overrightarrow{\mathbf{h}}_m]$. Each candidate mention
 8558 span (s, t) is then represented by the vertical concatenation of four vectors:

$$\mathbf{u}^{(s,t)} = [u_{\text{first}}^{(s,t)}; u_{\text{last}}^{(s,t)}; u_{\text{head}}^{(s,t)}; \phi^{(s,t)}], \quad [15.25]$$

8559 where $u_{\text{first}}^{(s,t)} = \mathbf{h}_{s+1}$ is the embedding of the first word in the span, $u_{\text{last}}^{(s,t)} = \mathbf{h}_t$ is the
 8560 embedding of the last word, $u_{\text{head}}^{(s,t)}$ is the embedding of the “head” word, and $\phi^{(s,t)}$ is a
 8561 vector of surface features, such as the length of the span (Lee et al., 2017).

Attention over head words Rather than identifying the head word from the output of a parser, it can be computed from a neural **attention mechanism**:

$$\tilde{\alpha}_m = \theta_\alpha \cdot \mathbf{h}_m \quad [15.26]$$

$$\mathbf{a}^{(s,t)} = \text{SoftMax}([\tilde{\alpha}_{s+1}, \tilde{\alpha}_{s+2}, \dots, \tilde{\alpha}_t]) \quad [15.27]$$

$$\mathbf{u}_{\text{head}}^{(s,t)} = \sum_{m=s+1}^t a_m^{(s,t)} \mathbf{h}_m. \quad [15.28]$$

8562 Each token m gets a scalar score $\tilde{\alpha}_m = \theta_\alpha \cdot \mathbf{h}_m$, which is the dot product of the LSTM
 8563 hidden state \mathbf{h}_m and a vector of weights θ_α . The vector of scores for tokens in the span
 8564 $m \in \{s + 1, s + 2, \dots, t\}$ is then passed through a softmax layer, yielding a vector $\mathbf{a}^{(s,t)}$
 8565 that allocates one unit of attention across the span. This eliminates the need for syntactic
 8566 parsing to recover the head word; instead, the model learns to identify the most important
 8567 words in each span. Attention mechanisms were introduced in neural machine transla-
 8568 tion (Bahdanau et al., 2014), and are described in more detail in § 18.3.1.

Using mention embeddings Given a set of mention embeddings, each mention i and candidate antecedent a is scored as,

$$\psi(a, i) = \psi_S(a) + \psi_S(i) + \psi_M(a, i) \quad [15.29]$$

$$\psi_S(a) = \text{FeedForward}_S(\mathbf{u}^{(a)}) \quad [15.30]$$

$$\psi_S(i) = \text{FeedForward}_S(\mathbf{u}^{(i)}) \quad [15.31]$$

$$\psi_M(a, i) = \text{FeedForward}_M([\mathbf{u}^{(a)}; \mathbf{u}^{(i)}; \mathbf{u}^{(a)} \odot \mathbf{u}^{(i)}; \mathbf{f}(a, i, \mathbf{w})]), \quad [15.32]$$

where $\mathbf{u}^{(a)}$ and $\mathbf{u}^{(i)}$ are the embeddings for spans a and i respectively, as defined in Equation 15.25.

- The scores $\psi_S(a)$ quantify whether span a is likely to be a coreferring mention, independent of what it corefers with. This allows the model to learn identify mentions directly, rather than identifying mentions with a preprocessing step.
- The score $\psi_M(a, i)$ computes the compatibility of spans a and i . Its base layer is a vector that includes the embeddings of spans a and i , their elementwise product $\mathbf{u}^{(a)} \odot \mathbf{u}^{(i)}$, and a vector of surface features $\mathbf{f}(a, i, \mathbf{w})$, including distance, speaker, and genre information.

Lee et al. (2017) provide an error analysis that shows how this method can correctly link a *blaze* and a *fire*, while incorrectly linking *pilots* and *fight attendants*. In each case, the coreference decision is based on similarities in the word embeddings.

Rather than embedding individual mentions, Clark and Manning (2016) embed mention pairs. At the base layer, their network takes embeddings of the words in and around each mention, as well as one-hot vectors representing a few surface features, such as the distance and string matching features. This base layer is then passed through a multilayer feedforward network with ReLU nonlinearities, resulting in a representation of the mention pair. The output of the mention pair encoder $\mathbf{u}_{i,j}$ is used in the scoring function of a mention-ranking model, $\psi_M(i, j) = \theta \cdot \mathbf{u}_{i,j}$. A similar approach is used to score cluster pairs, constructing a cluster-pair encoding by **pooling** over the mention-pair encodings for all pairs of mentions within the two clusters.

Entity embeddings

In entity-based coreference resolution, each entity should be represented by properties of its mentions. In a distributed setting, we maintain a set of vector entity embeddings, \mathbf{v}_e . Each candidate mention receives an embedding \mathbf{u}_i ; Wiseman et al. (2016) compute this embedding by a single-layer neural network, applied to a vector of surface features. The decision of whether to merge mention i with entity e can then be driven by a feedforward

network, $\psi_E(i, e) = \text{Feedforward}([\mathbf{v}_e; \mathbf{u}_i])$. If i is added to entity e , then its representation is updated recurrently, $\mathbf{v}_e \leftarrow f(\mathbf{v}_e, \mathbf{u}_i)$, using a recurrent neural network such as a long short-term memory (LSTM; chapter 6). Alternatively, we can apply a pooling operation, such as max-pooling or average-pooling (chapter 3), setting $\mathbf{v}_e \leftarrow \text{Pool}(\mathbf{v}_e, \mathbf{u}_i)$. In either case, the update to the representation of entity e can be thought of as adding new information about the entity from mention i .

15.4 Evaluating coreference resolution

The state of coreference evaluation is aggravatingly complex. Early attempts at simple evaluation metrics were found to be susceptible to trivial baselines, such as placing each mention in its own cluster, or grouping all mentions into a single cluster. Following Denis and Baldridge (2009), the CoNLL 2011 shared task on coreference (Pradhan et al., 2011) formalized the practice of averaging across three different metrics: MUC (Vilain et al., 1995), B-CUBED (Bagga and Baldwin, 1998a), and CEAf (Luo, 2005). Reference implementations of these metrics are available from Pradhan et al. (2014) at <https://github.com/conll/reference-coreference-scorers>.

Additional resources

Ng (2010) surveys coreference resolution through 2010. Early work focused exclusively on pronoun resolution, with rule-based (Lappin and Leass, 1994) and probabilistic methods (Ge et al., 1998). The full coreference resolution problem was popularized in a shared task associated with the sixth Message Understanding Conference, which included coreference annotations for training and test sets of thirty documents each (Grishman and Sundheim, 1996). An influential early paper was the decision tree approach of Soon et al. (2001), who introduced mention ranking. A comprehensive list of surface features for coreference resolution is offered by Bengtson and Roth (2008). Durrett and Klein (2013) improved on prior work by introducing a large lexicalized feature set; subsequent work has emphasized neural representations of entities and mentions (Wiseman et al., 2015).

Exercises

1. Select an article from today’s news, and annotate coreference for the first twenty noun phrases that appear in the article (include nested noun phrases). Then specify the mention-pair training data that would result from the first five noun phrases.
2. Using your annotations from the preceding problem, compute the following statistics:

- 8628 • The number of times new entities are introduced by each of the three types of
 8629 referring expressions: pronouns, proper nouns, and nominals. Include “single-
 8630 ton” entities that are mentioned only once.
 - 8631 • For each type of referring expression, compute the fraction of mentions that are
 8632 anaphoric.
- 8633 3. Apply a simple heuristic to all pronouns in the article from the previous exercise:
 8634 link each pronoun to the closest preceding noun phrase that agrees in gender, num-
 8635 ber, animacy, and person. Compute the following evaluation:
- 8636 • True positive: a pronoun that is linked to a noun phrase with which it is coref-
 8637 erent, or is labeled as the first mention of an entity when in fact it does not
 8638 corefer with any preceding mention. In this case, non-referential pronouns can
 8639 be true positives if they are marked as having no antecedent.
 - 8640 • False positive: a pronoun that is linked to a noun phrase with which it is not
 8641 coreferent. This includes mistakenly linking singleton or non-referential pro-
 8642 nouns.
 - 8643 • False negative: a pronoun that has at least one antecedent, but is either labeled
 8644 as not having an antecedent, or is linked to mention with which it does not
 8645 corefer.
- 8646 Compute the *F*-MEASURE for your method, and for a trivial baseline in which ev-
 8647 ery mention is its own entity. Are there any additional heuristics that would have
 8648 improved the performance of this method?
- 8649 4. Durrett and Klein (2013) compute the probability of the gold coreference clustering
 8650 by summing over all antecedent structures that are compatible with the clustering.
 8651 For example, if there are three mentions of a single entity, m_1, m_2, m_3 , there are two
 8652 possible antecedent structures: $a_2 = 1, a_3 = 1$ and $a_2 = 1, a_3 = 2$. Compute the
 8653 number of antecedent structures for a single entity with K mentions.
- 8654 5. Suppose that all mentions can be unambiguously divided into C classes, for exam-
 8655 ple by gender and number. Further suppose that mentions from different classes
 8656 can never corefer. In a document with M mentions, give upper and lower bounds
 8657 on the total number of possible coreference clusterings, in terms of the Bell numbers
 8658 and the parameters M and C . Compute numerical upper and lower bounds for the
 8659 case $M = 4, C = 2$.
- 8660 6. Lee et al. (2017) propose a model that considers all contiguous spans in a document
 8661 as possible mentions.
- 8662 a) In a document of length M , how many mention pairs must be evaluated? (All
 8663 answers can be given in asymptotic, big-O notation.)

- 8664 b) To make inference more efficient, Lee et al. (2017) restrict consideration to spans
 8665 of maximum length $L \ll M$. Under this restriction, how many mention pairs
 8666 must be evaluated?
 8667 c) To further improve inference, one might evaluate coreference only between
 8668 pairs of mentions whose endpoints are separated by a maximum of D tokens.
 8669 Under this additional restriction, how many mention pairs must be evaluated?

8670 7. In Spanish, the subject can be omitted when it is clear from context, e.g.,

- 8671 (15.13) *Las ballenas no son peces. Son mamíferos.*
 8672 The whales no are fish. Are mammals.
 8672 Whales are not fish. They are mammals.

8673 Resolution of such **null subjects** is facilitated by the Spanish system of verb mor-
 8674 phology, which includes distinctive suffixes for most combinations of person and
 8675 number. For example, the verb form *son* ('are') agrees with the third-person plural
 8676 pronouns *ellos* (masculine) and *ellas* (feminine), as well as the second-person plural
 8677 *ustedes*.

8678 Suppose that you are given the following components:

- 8679 • A system that automatically identifies verbs with null subjects.
- 8680 • A function $c(j, p) \in \{0, 1\}$ that indicates whether pronoun p is compatible with
 8681 null subject j , according to the verb morphology.
- 8682 • A trained mention-pair model, which computes scores $\psi(w_i, w_j, j - i) \in \mathbb{R}$ for
 8683 all pairs of mentions i and j , scoring the pair by the antecedent mention w_i , the
 8684 anaphor w_j , and the distance $j - i$.

8685 Describe an integer linear program that simultaneously performs two tasks: resolv-
 8686 ing coreference among all entity mentions, and identifying suitable pronouns for all
 8687 null subjects. In the example above, your program should link the null subject with
 8688 *las ballenas* ('whales'), and identify *ellas* as the correct pronoun. For simplicity, you
 8689 may assume that null subjects cannot be antecedents, and you need not worry about
 8690 the transitivity constraint described in § 15.2.3.

8691 8. Use the policy gradient algorithm to compute the gradient for the following sce-
 8692 nario, based on the Bell tree in Figure 15.4:

- 8693 • The gold clustering c^* is $\{Abigail, her\}, \{she\}$.

- Drawing a single sequence of actions ($K = 1$) from the current policy, you obtain the following incremental clusterings:

$$\begin{aligned}\mathbf{c}(a_1) &= \{\text{Abigail}\} \\ \mathbf{c}(a_{1:2}) &= \{\text{Abigail}, \text{she}\} \\ \mathbf{c}(a_{1:3}) &= \{\text{Abigail}, \text{she}\}, \{\text{her}\}.\end{aligned}$$

- 8694 • At each mention t , the space of actions \mathcal{A}_t includes merging the mention with
 8695 each existing cluster or with the empty cluster. The probability of merging m_t
 8696 with cluster c is proportional to the exponentiated score for the merged cluster,

$$p(\text{Merge}(m_t, c)) \propto \exp \psi_E(m_t \cup c), \quad [15.33]$$

8697 where $\psi_E(m_t \cup c)$ is defined in Equation 15.15.

8698 Compute the gradient $\frac{\partial}{\partial \theta} L(\theta)$ in terms of the loss $\ell(\mathbf{c}(\mathbf{a}))$ and the features of each
 8699 (potential) cluster. Explain the differences between the gradient-based update $\theta \leftarrow \theta - \frac{\partial}{\partial \theta} L(\theta)$
 8700 and the incremental perceptron update from this same example.

- 8701 9. As discussed in § 15.1.1, some pronouns are not referential. In English, this occurs
 8702 frequently with the word *it*. Download the text of *Alice in Wonderland* from NLTK,
 8703 and examine the first ten appearances of *it*. For each occurrence:

- 8704 • First, examine a five-token window around the word. In the first example, this
 8705 window is,

8706 , but it had no

8707 Is there another pronoun that could be substituted for *it*? Consider *she*, *they*,
 8708 and *them*. In this case, both *she* and *they* yield grammatical substitutions. What
 8709 about the other ten appearances of *it*?

- 8710 • Now, view an fifteen-word window for each example. Based on this window,
 8711 mark whether you think the word *it* is referential.

8712 How often does the substitution test predict whether *it* is referential?

- 8713 10. Now try to automate the test, using the Google n -grams corpus (Brants and Franz,
 8714 2006). Specifically, find the count of each 5-gram containing *it*, and then compute
 8715 the counts of 5-grams in which *it* is replaced with other third-person pronouns: *he*,
 8716 *she*, *they*, *her*, *him*, *them*, *herself*, *himself*.

8717 There are various ways to get these counts. One approach is to download the
 8718 raw data and search it; another is to construct web queries to <https://books.google.com/ngrams>.

8720 Compare the ratio of the counts of the original 5-gram to the summed counts of
8721 the 5-grams created by substitution. Is this ratio a good predictor of whether *it* is
8722 referential?

8723 Chapter 16

8724 Discourse

8725 Applications of natural language processing often concern multi-sentence documents:
8726 from paragraph-long restaurant reviews, to 500-word newspaper articles, to 500-page
8727 novels. Yet most of the methods that we have discussed thus far are concerned with
8728 individual sentences. This chapter discusses theories and methods for handling multi-
8729 sentence linguistic phenomena, known collectively as **discourse**. There are diverse char-
8730 acterizations of discourse structure, and no single structure is ideal for every computa-
8731 tional application. This chapter covers some of the most well studied discourse repre-
8732 sentations, while highlighting computational models for identifying and exploiting these
8733 structures.

8734 16.1 Segments

8735 A document or conversation can be viewed as a sequence of **segments**, each of which is
8736 **cohesive** in its content and/or function. In Wikipedia biographies, these segments often
8737 pertain to various aspects to the subject's life: early years, major events, impact on others,
8738 and so on. This segmentation is organized around **topics**. Alternatively, scientific research
8739 articles are often organized by **functional themes**: the introduction, a survey of previous
8740 research, experimental setup, and results.

8741 Written texts often mark segments with section headers and related formatting de-
8742 vices. However, such formatting may be too coarse-grained to support applications such
8743 as the retrieval of specific passages of text that are relevant to a query (Hearst, 1997).
8744 Unformatted speech transcripts, such as meetings and lectures, are also an application
8745 scenario for segmentation (Carletta, 2007; Glass et al., 2007; Janin et al., 2003).

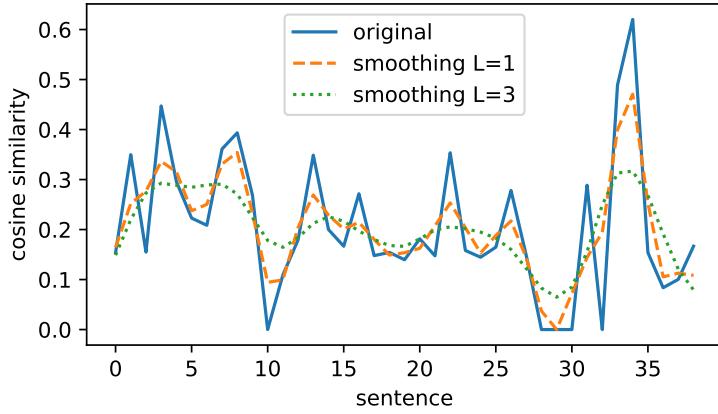


Figure 16.1: Smoothed cosine similarity among adjacent sentences in a news article. Local minima at $m = 10$ and $m = 29$ indicate likely segmentation points.

8746 16.1.1 Topic segmentation

A cohesive topic segment forms a unified whole, using various linguistic devices: repeated references to an entity or event; the use of conjunctions to link related ideas; and the repetition of meaning through lexical choices (Halliday and Hasan, 1976). Each of these cohesive devices can be measured, and then used as features for topic segmentation. A classical example is the use of lexical cohesion in the TEXTTILING method for topic segmentation (Hearst, 1997). The basic idea is to compute the textual similarity between each pair of adjacent blocks of text (sentences or fixed-length units), using a formula such as the smoothed **cosine similarity** of their bag-of-words vectors,

$$s_m = \frac{\mathbf{x}_m \cdot \mathbf{x}_{m+1}}{\|\mathbf{x}_m\|_2 \times \|\mathbf{x}_{m+1}\|_2} \quad [16.1]$$

$$\bar{s}_m = \sum_{\ell=0}^L k_\ell (s_{m+\ell} + s_{m-\ell}), \quad [16.2]$$

8747 with k_ℓ representing the value of a smoothing kernel of size L , e.g. $\mathbf{k} = [1, 0.5, 0.25]^\top$.
 8748 Segmentation points are then identified at local minima in the smoothed similarities \bar{s} ,
 8749 since these points indicate changes in the overall distribution of words in the text. An
 8750 example is shown in Figure 16.1.

8751 Text segmentation can also be formulated as a probabilistic model, in which each seg-
 8752 ment has a unique language model that defines the probability over the text in the seg-
 8753 ment (Utiyama and Isahara, 2001; Eisenstein and Barzilay, 2008; Du et al., 2013).¹ A good

¹There is a rich literature on how latent variable models (such as **latent Dirichlet allocation**) can track

8754 segmentation achieves high likelihood by grouping segments with similar word distribu-
8755 tions. This probabilistic approach can be extended to **hierarchical topic segmentation**, in
8756 which each topic segment is divided into subsegments (Eisenstein, 2009). All of these ap-
8757 proaches are unsupervised. While labeled data can be obtained from well-formatted texts
8758 such as textbooks, such annotations may not generalize to speech transcripts in alterna-
8759 tive domains. Supervised methods have been tried in cases where in-domain labeled data
8760 is available, substantially improving performance by learning weights on multiple types
8761 of features (Galley et al., 2003).

8762 16.1.2 Functional segmentation

8763 In some genres, there is a canonical set of communicative *functions*: for example, in sci-
8764 entific research articles, one such function is to communicate the general background for
8765 the article, another is to introduce a new contribution, or to describe the aim of the re-
8766 search (Teufel et al., 1999). A **functional segmentation** divides the document into con-
8767 tiguous segments, sometimes called **rhetorical zones**, in which each sentence has the same
8768 function. Teufel and Moens (2002) train a supervised classifier to identify the functional
8769 of each sentence in a set of scientific research articles, using features that describe the sen-
8770 tence's position in the text, its similarity to the rest of the article and title, tense and voice of
8771 the main verb, and the functional role of the previous sentence. Functional segmentation
8772 can also be performed without supervision. Noting that some types of Wikipedia arti-
8773 cles have very consistent functional segmentations (e.g., articles about cities or chemical
8774 elements), Chen et al. (2009) introduce an unsupervised model for functional segmenta-
8775 tion, which learns both the language model associated with each function and the typical
8776 patterning of functional segments across the article.

8777 16.2 Entities and reference

8778 Another dimension of discourse relates to which entities are mentioned throughout the
8779 text, and how. Consider the examples in Figure 16.2: Grosz et al. (1995) argue that the first
8780 discourse is more coherent. Do you agree? The examples differ in their choice of **refe-
8781 ring expressions** for the protagonist *John*, and in the syntactic constructions in sentences
8782 (b) and (d). The examples demonstrate the need for theoretical models to explain how
8783 referring expressions are chosen, and where they are placed within sentences. Such mod-
8784 els can then be used to help interpret the overall structure of the discourse, to measure
8785 discourse coherence, and to generate discourses in which referring expressions are used
8786 coherently.

topics across documents (Blei et al., 2003; Blei, 2012).

- | | |
|--|---|
| (16.1) a. John went to his favorite music store to buy a piano.
b. He had frequented the store for many years.
c. He was excited that he could finally buy a piano.
d. He arrived just as the store was closing for the day | (16.2) a. John went to his favorite music store to buy a piano.
b. It was a store John had frequented for many years.
c. He was excited that he could finally buy a piano.
d. It was closing just as John arrived. |
|--|---|

Figure 16.2: Two tellings of the same story (Grosz et al., 1995). The discourse on the left uses referring expressions coherently, while the one on the right does not.

8787 16.2.1 Centering theory

8788 **Centering theory** presents a unified account of the relationship between discourse struc-
 8789 ture and entity reference (Grosz et al., 1995). According to the theory, every utterance in
 8790 the discourse is characterized by a set of entities, known as *centers*.

- 8791 • The **forward-looking centers** in utterance m are all the entities that are mentioned
 8792 in the utterance, $c_f(w_m) = \{e_1, e_2, \dots\}$. The forward-looking centers are partially
 8793 ordered by their syntactic prominence, favoring subjects over objects, and objects
 8794 over other positions (Brennan et al., 1987). For example, in example (1.1a) of Fig-
 8795 ure 16.2, the ordered list of forward-looking centers in the first utterance is John, the
 8796 music store, and the piano.
- 8797 • The **backward-looking center** $c_b(w_m)$ is the highest-ranked element in the set of
 8798 forward-looking centers from the previous utterance $c_f(w_{m-1})$ that is also men-
 8799 tioned in w_m . In example (1.1b) of item 16.1, the backward looking center is John.

8800 Given these two definitions, centering theory makes the following predictions about
 8801 the form and position of referring expressions:

- 8802 1. If a pronoun appears in the utterance w_m , then the backward-looking center $c_b(w_m)$
 8803 must also be realized as a pronoun. This rule argues against the use of *it* to refer
 8804 to the piano store in Example (16.2d), since JOHN is the backward looking center of
 8805 (16.2d), and he is mentioned by name and not by a pronoun.
- 8806 2. Sequences of utterances should retain the same backward-looking center if possible,
 8807 and ideally, the backward-looking center should also be the top-ranked element in
 8808 the list of forward-looking centers. This rule argues in favor of the preservation of
 8809 JOHN as the backward-looking center throughout Example (16.1).

	SKYLER	WALTER	DANGER	A GUY	THE DOOR
<i>You don't know who you're talking to,</i>	S	-	-	-	-
<i>so let me clue you in.</i>	O	O	-	-	-
<i>I am not in danger, Skyler.</i>	X	S	X	-	-
<i>I am the danger.</i>	-	S	O	-	-
<i>A guy opens his door and gets shot,</i>	-	-	-	S	O
<i>and you think that of me?</i>	S	X	-	-	-
<i>No. I am the one who knocks!</i>	-	S	-	-	-

Figure 16.3: The entity grid representation for a dialogue from the television show *Breaking Bad*.

8810 Centering theory unifies aspects of syntax, discourse, and anaphora resolution. However,
 8811 it can be difficult to clarify exactly how to rank the elements of each utterance, or even
 8812 how to partition a text or dialog into utterances (Poesio et al., 2004).

8813 16.2.2 The entity grid

8814 One way to formalize the ideas of centering theory is to arrange the entities in a text or
 8815 conversation in an **entity grid**. This is a data structure with one row per sentence, and
 8816 one column per entity (Barzilay and Lapata, 2008). Each cell $c(m, i)$ can take the following
 8817 values:

$$c(m, i) = \begin{cases} S, & \text{entity } i \text{ is in subject position in sentence } m \\ O, & \text{entity } i \text{ is in object position in sentence } m \\ X, & \text{entity } i \text{ appears in sentence } m, \text{ in neither subject nor object position} \\ -, & \text{entity } i \text{ does not appear in sentence } m. \end{cases} \quad [16.3]$$

8818 To populate the entity grid, syntactic parsing is applied to identify subject and object
 8819 positions, and coreference resolution is applied to link multiple mentions of a single entity.
 8820 An example is shown in Figure 16.3.

8821 After the grid is constructed, the coherence of a document can be measured by the
 8822 transitions between adjacent cells in each column. For example, the transition $(S \rightarrow S)$
 8823 keeps an entity in subject position across adjacent sentences; the transition $(O \rightarrow S)$ pro-
 8824 motes an entity from object position to subject position; the transition $(S \rightarrow -)$ drops the
 8825 subject of one sentence from the next sentence. The probabilities of each transition can be

estimated from labeled data, and an entity grid can then be scored by the sum of the log-probabilities across all columns and all transitions, $\sum_{i=1}^{N_e} \sum_{m=1}^M \log p(c(m, i) | c(m - 1, i))$. The resulting probability can be used as a proxy for the coherence of a text. This has been shown to be useful for a range of tasks: determining which of a pair of articles is more readable (Schwartz and Ostendorf, 2005), correctly ordering the sentences in a scrambled text (Lapata, 2003), and disentangling multiple conversational threads in an online multi-party chat (Elsner and Charniak, 2010).

16.2.3 *Formal semantics beyond the sentence level

An alternative view of the role of entities in discourse focuses on formal semantics, and the construction of meaning representations for multi-sentence units. Consider the following two sentences (from Bird et al., 2009):

- (16.3) a. Angus owns a dog.
- b. It bit Irene.

We would like to recover the formal semantic representation,

$$\exists x. \text{DOG}(x) \wedge \text{OWN}(\text{ANGUS}, x) \wedge \text{BITE}(x, \text{IRENE}). \quad [16.4]$$

However, the semantic representations of each individual sentence are,

$$\exists x. \text{DOG}(x) \wedge \text{OWN}(\text{ANGUS}, x) \quad [16.5]$$

$$\text{BITE}(y, \text{IRENE}). \quad [16.6]$$

Unifying these two representations into the form of Equation 16.4 requires linking the unbound variable y from [16.6] with the quantified variable x in [16.5].² Discourse understanding therefore requires the reader to update a set of assignments, from variables to entities. This update would (presumably) link the *dog* in the first sentence of [16.3] with the unbound variable y in the second sentence, thereby licensing the conjunction in [16.4].³ This basic idea is at the root of **dynamic semantics** (Groenendijk and Stokhof, 1991). **Segmented discourse representation theory** links dynamic semantics with a set of **discourse relations**, which explain how adjacent units of text are rhetorically or conceptually related (Lascarides and Asher, 2007). The next section explores the theory of discourse relations in more detail.

²Groenendijk and Stokhof (1991) treats the y variable in Equation 16.6 as unbound. Even if it were bound locally with an existential quantifier ($\exists y \text{BITE}(y, \text{IRENE})$), the variable would still need to be reconciled with the quantified variable in Equation 16.5.

³This linking task is similar to coreference resolution (see chapter 15), but here the connections are between semantic variables, rather than spans of text.

8850 16.3 Relations

8851 In dependency grammar, sentences are characterized by a graph (usually a tree) of syntac-
8852 tic relations between words, such as NSUBJ and DET. A similar idea can be applied at the
8853 document level, identifying relations between discourse units, such as clauses, sentences,
8854 or paragraphs. The task of **discourse parsing** involves identifying discourse units and
8855 the relations that hold between them. These relations can then be applied to tasks such as
8856 document classification and summarization, as discussed in § 16.3.4.

8857 16.3.1 Shallow discourse relations

8858 The existence of discourse relations is hinted by **discourse connectives**, such as *however*,
8859 *moreover*, *meanwhile*, and *if ... then*. These connectives explicitly specify the relationship
8860 between adjacent units of text: *however* signals a contrastive relationship, *moreover* signals
8861 that the subsequent text elaborates or strengthens the point that was made immediately
8862 beforehand, *meanwhile* indicates that two events are contemporaneous, and *if ... then* sets
8863 up a conditional relationship. Discourse connectives can therefore be viewed as a starting
8864 point for the analysis of discourse relations.

8865 In **lexicalized tree-adjoining grammar for discourse (D-LTAG)**, each connective an-
8866 chors a relationship between two units of text (Webber, 2004). This model provides the
8867 theoretical basis for the **Penn Discourse Treebank (PDTB)**, the largest corpus of discourse
8868 relations in English (Prasad et al., 2008). It includes a hierarchical inventory of discourse
8869 relations (shown in Table 16.1), which is created by abstracting the meanings implied by
8870 the discourse connectives that appear in real texts (Knott, 1996). These relations are then
8871 annotated on the same corpus of news text used in the Penn Treebank (see § 9.2.2), adding
8872 the following information:

- 8873 • Each connective is annotated for the discourse relation or relations that it expresses,
8874 if any — many discourse connectives have senses in which they do not signal a
8875 discourse relation (Pitler and Nenkova, 2009).
- 8876 • For each discourse relation, the two arguments of the relation are specified as ARG1
8877 and ARG2, where ARG2 is constrained to be adjacent to the connective. These argu-
8878 ments may be sentences, but they may also smaller or larger units of text.
- 8879 • Adjacent sentences are annotated for **implicit discourse relations**, which are not
8880 marked by any connective. When a connective could be inserted between a pair
8881 of sentence, the annotator supplies it, and also labels its sense (e.g., example 16.5).
8882 In some cases, there is no relationship at all between a pair of adjacent sentences;
8883 in other cases, the only relation is that the adjacent sentences mention one or more
8884 shared entity. These phenomena are annotated as NOREL and ENTREL (entity rela-
8885 tion), respectively.

- TEMPORAL
 - Asynchronous
 - Synchronous: precedence, succession
- CONTINGENCY
 - Cause: result, reason
 - Pragmatic cause: justification
 - Condition: hypothetical, general, unreal present, unreal past, real present, real past
 - Pragmatic condition: relevance, implicit assertion
- COMPARISON
 - Contrast: juxtaposition, opposition
 - Pragmatic contrast
 - Concession: expectation, contra-expectation
 - Pragmatic concession
- EXPANSION
 - Conjunction
 - Instantiation
 - Restatement: specification, equivalence, generalization
 - Alternative: conjunctive, disjunctive, chosen alternative
 - Exception
 - List

Table 16.1: The hierarchy of discourse relation in the Penn Discourse Treebank annotations (Prasad et al., 2008). For example, PRECEDENCE is a subtype of SYNCHRONOUS, which is a type of TEMPORAL relation.

886 Examples of Penn Discourse Treebank annotations are shown in (16.4). In (16.4), the
 887 word *therefore* acts as an explicit discourse connective, linking the two adjacent units of
 888 text. The Treebank annotations also specify the “sense” of each relation, linking the con-
 889 nective to a relation in the sense inventory shown in Table 16.1: in (16.4), the relation is
 890 PRAGMATIC CAUSE:JUSTIFICATION because it relates to the author’s communicative in-
 891 tentions. The word *therefore* can also signal causes in the external world (e.g., *He was*
 892 *therefore forced to relinquish his plan*). In **discourse sense classification**, the goal is to de-
 893 termine which discourse relation, if any, is expressed by each connective. A related task
 894 is the classification of implicit discourse relations, as in (16.5). In this example, the re-
 895 lationship between the adjacent sentences could be expressed by the connective *because*,
 896 indicating a CAUSE:REASON relationship.

887 **Classifying explicit discourse relations and their arguments**

888 As suggested by the examples above, many connectives can be used to invoke multiple
 889 types of discourse relations. Similarly, some connectives have senses that are unrelated
 890 to discourse: for example, *and* functions as a discourse connective when it links propo-

- (16.4) *...as this business of whaling has somehow come to be regarded among landsmen as a rather unpoetical and disreputable pursuit; therefore, I am all anxiety to convince ye, ye landsmen, of the injustice hereby done to us hunters of whales.*
- (16.5) But a few funds have taken other defensive steps. *Some have raised their cash positions to record levels. Implicit = BECAUSE High cash positions help buffer a fund when the market falls.*
- (16.6) Michelle lives in a hotel room, and although **she drives a canary-colored Porsche**, *she hasn't time to clean or repair it.*
- (16.7) *Most oil companies, when they set exploration and production budgets for this year, forecast revenue of \$15 for each barrel of crude produced.*

Figure 16.4: Example annotations of discourse relations. In the style of the Penn Discourse Treebank, the discourse connective is underlined, the first argument is shown in italics, and the second argument is shown in bold. Examples (16.5-16.7) are quoted from Prasad et al. (2008).

8901 sitions, but not when it links noun phrases (Lin et al., 2014). Nonetheless, the senses of
 8902 explicitly-marked discourse relations in the Penn Treebank are relatively easy to classify,
 8903 at least at the coarse-grained level. When classifying the four top-level PDTB relations,
 8904 90% accuracy can be obtained simply by selecting the most common relation for each
 8905 connective (Pitler and Nenkova, 2009). At the more fine-grained levels of the discourse
 8906 relation hierarchy, connectives are more ambiguous. This fact is reflected both in the ac-
 8907 curacy of automatic sense classification (Versley, 2011) and in interannotator agreement,
 8908 which falls to 80% for level-3 discourse relations (Prasad et al., 2008).

8909 A more challenging task for explicitly-marked discourse relations is to identify the
 8910 scope of the arguments. Discourse connectives need not be adjacent to ARG1, as shown
 8911 in item 16.6, where ARG1 follows ARG2; furthermore, the arguments need not be contigu-
 8912 ous, as shown in (16.7). For these reasons, recovering the arguments of each discourse
 8913 connective is a challenging subtask. Because intra-sentential arguments are often syn-
 8914 tactic constituents (see chapter 10), many approaches train a classifier to predict whether
 8915 each constituent is an appropriate argument for each explicit discourse connective (Well-
 8916 ner and Pustejovsky, 2007; Lin et al., 2014, e.g.).

8917 Classifying implicit discourse relations

Implicit discourse relations are considerably more difficult to classify and to annotate.⁴
 Most approaches are based on an encoding of each argument, which is then used as input

⁴In the dataset for the 2015 shared task on shallow discourse parsing, the interannotator agreement was 91% for explicit discourse relations and 81% for implicit relations, across all levels of detail (Xue et al., 2015).

to a nonlinear classifier:

$$\mathbf{z}^{(i)} = \text{Encode}(\mathbf{w}^{(i)}) \quad [16.7]$$

$$\mathbf{z}^{(i+1)} = \text{Encode}(\mathbf{w}^{(i+1)}) \quad [16.8]$$

$$\hat{y}_i = \underset{y}{\operatorname{argmax}} \Psi(y, \mathbf{z}^{(i)}, \mathbf{z}^{(i+1)}). \quad [16.9]$$

- 8918 This basic framework can be instantiated in several ways, including both feature-based
 8919 and neural encoders.

8920 **Feature-based approaches** Each argument can be encoded into a vector of surface fea-
 8921 tures. The encoding typically includes lexical features (all words, or all content words, or
 8922 a subset of words such as the first three and the main verb), Brown clusters of individ-
 8923 ual words (§ 14.4), and syntactic features such as terminal productions and dependency
 8924 arcs (Pitler et al., 2009; Lin et al., 2009; Rutherford and Xue, 2014). The classification func-
 8925 tion then has two parts. First, it creates a joint feature vector by combining the encodings
 8926 of each argument, typically by computing the cross-product of all features in each encod-
 8927 ing:

$$\mathbf{f}(y, \mathbf{z}^{(i)}, \mathbf{z}^{(i+1)}) = \{(a \times b \times y) : (\mathbf{z}_a^{(i)} \mathbf{z}_b^{(i+1)})\} \quad [16.10]$$

8928 The size of this feature set grows with the square of the size of the vocabulary, so it can be
 8929 helpful to select a subset of features that are especially useful on the training data (Park
 8930 and Cardie, 2012). After \mathbf{f} is computed, any classifier can be trained to compute the final
 8931 score, $\Psi(y, \mathbf{z}^{(i)}, \mathbf{z}^{(i+1)}) = \theta \cdot \mathbf{f}(y, \mathbf{z}^{(i)}, \mathbf{z}^{(i+1)})$.

8932 **Neural network approaches** In neural network architectures, the encoder is learned
 8933 jointly with the classifier as an end-to-end model. Each argument can be encoded using
 8934 a variety of neural architectures (surveyed in § 14.8): recursive (§ 10.6.1; Ji and Eisenstein,
 8935 2015), recurrent (§ 6.3; Ji et al., 2016), and convolutional (§ 3.4; Qin et al., 2017). The clas-
 8936 sification function can then be implemented as a feedforward neural network on the two
 8937 encodings (chapter 3; for examples, see Rutherford et al., 2017; Qin et al., 2017), or as a
 8938 simple bilinear product, $\Psi(y, \mathbf{z}^{(i)}, \mathbf{z}^{(i+1)}) = (\mathbf{z}^{(i)})^\top \Theta_y \mathbf{z}^{(i+1)}$ (Ji and Eisenstein, 2015). The
 8939 encoding model can be trained by backpropagation from the classification objective, such
 8940 as the margin loss. Rutherford et al. (2017) show that neural architectures outperform
 8941 feature-based approaches in most settings. While neural approaches require engineering
 8942 the network architecture (e.g., embedding size, number of hidden units in the classifier),
 8943 feature-based approaches also require significant engineering to incorporate linguistic re-
 8944 sources such as Brown clusters and parse trees, and to select a subset of relevant features.

8945 **16.3.2 Hierarchical discourse relations**

8946 In sentence parsing, adjacent phrases combine into larger constituents, ultimately pro-
 8947 ducing a single constituent for the entire sentence. The resulting tree structure enables
 8948 structured analysis of the sentence, with subtrees that represent syntactically coherent
 8949 chunks of meaning. **Rhetorical Structure Theory (RST)** extends this style of hierarchical
 8950 analysis to the discourse level (Mann and Thompson, 1988).

8951 The basic element of RST is the **discourse unit**, which refers to a contiguous span of
 8952 text. **Elementary discourse units** (EDUs) are the atomic elements in this framework, and
 8953 are typically (but not always) clauses.⁵ Each discourse relation combines two or more
 8954 adjacent discourse units into a larger, composite discourse unit; this process ultimately
 8955 unites the entire text into a tree-like structure.⁶

8956 **Nuclearity** In many discourse relations, one argument is primary. For example:

- 8957 (16.8) [LaShawn loves animals]_{*N*}
 8958 [She has nine dogs and one pig]_{*S*}

8959 In this example, the second sentence provides EVIDENCE for the point made in the first
 8960 sentence. The first sentence is thus the **nucleus** of the discourse relation, and the second
 8961 sentence is the **satellite**. The notion of nuclearity is similar to the head-modifier structure
 8962 of dependency parsing (see § 11.1.1). However, in RST, some relations have multiple
 8963 nuclei. For example, the arguments of the CONTRAST relation are equally important:

- 8964 (16.9) [The clash of ideologies survives this treatment]_{*N*}
 8965 [but the nuance and richness of Gorky's individual characters have vanished in the scuffle]_{*N*}⁷

8966 Relations that have multiple nuclei are called **coordinating**; relations with a single nu-
 8967 cleus are called **subordinating**. Subordinating relations are constrained to have only two
 8968 arguments, while coordinating relations (such as CONJUNCTION) may have more than
 8969 two.

⁵Details of discourse segmentation can be found in the RST annotation manual (Carlson and Marcu, 2001).

⁶While RST analyses are typically trees, this should not be taken as a strong theoretical commitment to the principle that all coherent discourses have a tree structure. Taboada and Mann (2006) write:

It is simply the case that trees are convenient, easy to represent, and easy to understand. There is, on the other hand, no theoretical reason to assume that trees are the only possible representation of discourse structure and of coherence relations.

The appropriateness of tree structures to discourse has been challenged, e.g., by Wolf and Gibson (2005), who propose a more general graph-structured representation.

⁷from the RST Treebank (Carlson et al., 2002)

8970 **RST Relations** Rhetorical structure theory features a large inventory of discourse relations, which are divided into two high-level groups: subject matter relations, and presentational relations. Presentational relations are organized around the intended beliefs of the reader. For example, in (16.8), the second discourse unit provides evidence intended to increase the reader’s belief in the proposition expressed by the first discourse unit, that *LaShawn loves animals*. In contrast, subject-matter relations are meant to communicate additional facts about the propositions contained in the discourse units that they relate:

8977 (16.10) [the debt plan was rushed to completion]_N
 8978 [in order to be announced at the meeting]_S⁸

8979 In this example, the satellite describes a world state that is realized by the action described
 8980 in the nucleus. This relationship is about the world, and not about the author’s communicative intentions.

8982 **Example** Figure 16.5 depicts an RST analysis of a paragraph from a movie review. Asymmetric (subordinating) relations are depicted with an arrow from the satellite to the nucleus; symmetric (coordinating) relations are depicted with lines. The elementary discourse units 1F and 1G are combined into a larger discourse unit with the symmetric CONJUNCTION relation. The resulting discourse unit is then the satellite in a JUSTIFY relation with 1E.

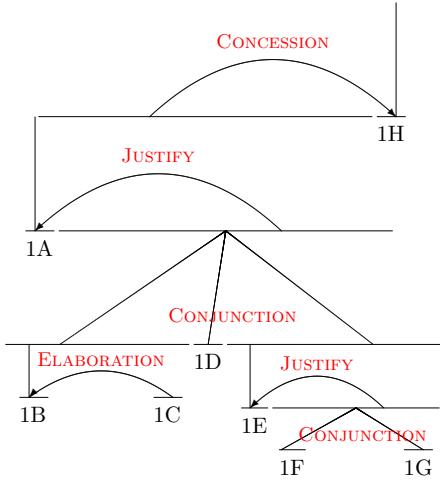
8988 Hierarchical discourse parsing

8989 The goal of discourse parsing is to recover a hierarchical structural analysis from a document text, such as the analysis in Figure 16.5. For now, let’s assume a segmentation of
 8990 the document into elementary discourse units (EDUs); segmentation algorithms are dis-
 8991 cussed below. After segmentation, discourse parsing can be viewed as a combination of
 8992 two components: the discourse relation classification techniques discussed in § 16.3.1, and
 8993 algorithms for phrase-structure parsing, such as chart parsing and shift-reduce, which
 8994 were discussed in chapter 10.

8996 Both chart parsing and shift-reduce require encoding composite discourse units, ei-
 8997 ther in a discrete feature vector or a dense neural representation.⁹ Some discourse parsers
 8998 rely on the **strong compositionality criterion** (Marcu, 1996), which states the assumption
 8999 that a composite discourse unit can be represented by its nucleus. This criterion is used in
 9000 feature-based discourse parsing to determine the feature vector for a composite discourse
 9001 unit (Hernault et al., 2010); it is used in neural approaches to setting the vector encod-
 9002 ing for a composite discourse unit equal to the encoding of its nucleus (Ji and Eisenstein,

⁸from the RST Treebank (Carlson et al., 2002)

⁹To use these algorithms, is also necessary to binarize all discourse relations during parsing, and then to “unbinarize” them to reconstruct the desired structure (e.g., Hernault et al., 2010).



[It could have been a great movie]^{1A} [It does have beautiful scenery,]^{1B} [some of the best since Lord of the Rings.]^{1C} [The acting is well done.]^{1D} [and I really liked the son of the leader of the Samurai.]^{1E} [He was a likable chap.]^{1F} [and I hated to see him die.]^{1G} [But, other than all that, this movie is nothing more than hidden rip-offs.]^{1H}

Figure 16.5: A rhetorical structure theory analysis of a short movie review, adapted from Voll and Taboada (2007). Positive and negative sentiment words are underlined, indicating RST’s potential utility in document-level sentiment analysis.

9003 2014). An alternative neural approach is to learn a composition function over the compo-
 9004 nents of a composite discourse unit (Li et al., 2014), using a recursive neural network (see
 9005 § 14.8.3).

9006 **Bottom-up discourse parsing** Assume a segmentation of the text into N elementary
 9007 discourse units with base representations $\{z^{(i)}\}_{i=1}^N$, and assume a composition function
 9008 $\text{COMPOSE}(z^{(i)}, z^{(j)}, \ell)$, which maps two encodings and a discourse relation ℓ into a new
 9009 encoding. The composition function can follow the strong compositionality criterion and
 9010 simply select the encoding of the nucleus, or it can do something more complex. We
 9011 also need a scoring function $\Psi(z^{(i,k)}, z^{(k,j)}, \ell)$, which computes a scalar score for the (bi-
 9012 narized) discourse relation ℓ with left child covering the span $i + 1 : k$, and the right
 9013 child covering the span $k + 1 : j$. Given these components, we can construct vector rep-
 9014 resentations for each span, and this is the basic idea underlying **compositional vector**
 9015 **grammars** (Socher et al., 2013).

9016 These same components can also be used in bottom-up parsing, in a manner that is
 9017 similar to the CKY algorithm for weighted context-free grammars (see § 10.1): compute
 9018 the score and best analysis for each possible span of increasing lengths, while storing

9019 back-pointers that make it possible to recover the optimal parse of the entire input. How-
 9020 ever, there is an important distinction from CKY parsing: for each labeled span (i, j, ℓ) , we
 9021 must use the composition function to construct a representation $z^{(i,j,\ell)}$. This representa-
 9022 tion is then used to combine the discourse unit spanning $i + 1 : j$ in higher-level discourse
 9023 relations. The representation $z^{(i,j,\ell)}$ depends on the entire substructure of the unit span-
 9024 ning $i + 1 : j$, and this violates the locality assumption that underlie CKY’s optimality
 9025 guarantee. Bottom-up parsing with recursively constructed span representations is gen-
 9026 erally not guaranteed to find the best-scoring discourse parse. This problem is explored
 9027 in an exercise at the end of the chapter.

9028 **Transition-based discourse parsing** One drawback of bottom-up parsing is its cubic
 9029 time complexity in the length of the input. For long documents, transition-based parsing
 9030 is an appealing alternative. The shift-reduce algorithm (see § 10.6.2) can be applied to
 9031 discourse parsing fairly directly (Sagae, 2009): the stack stores a set of discourse units and
 9032 their representations, and each action is chosen by a function of these representations.
 9033 This function could be a linear product of weights and features, or it could be a neural
 9034 network applied to encodings of the discourse units. The REDUCE action then performs
 9035 composition on the two discourse units at the top of the stack, yielding a larger composite
 9036 discourse unit, which goes on top of the stack. All of the techniques for integrating learn-
 9037 ing and transition-based parsing, described in § 11.3, are applicable to discourse parsing.

9038 Segmenting discourse units

9039 In rhetorical structure theory, elementary discourse units do not cross the sentence bound-
 9040 ary, so discourse segmentation can be performed within sentences, assuming the sentence
 9041 segmentation is given. The segmentation of sentences into elementary discourse units is
 9042 typically performed using features of the syntactic analysis (Braud et al., 2017). One ap-
 9043 proach is to train a classifier to determine whether each syntactic constituent is an EDU,
 9044 using features such as the production, tree structure, and head words (Soricut and Marcu,
 9045 2003; Hernault et al., 2010). Another approach is to train a sequence labeling model, such
 9046 as a conditional random field (Sporleder and Lapata, 2005; Xuan Bach et al., 2012; Feng
 9047 et al., 2014). This is done using the BIO formalism for segmentation by sequence labeling,
 9048 described in § 8.3.

9049 16.3.3 Argumentation

9050 An alternative view of text-level relational structure focuses on **argumentation** (Stab and
 9051 Gurevych, 2014b). Each segment (typically a sentence or clause) may support or rebut
 9052 another segment, creating a graph structure over the text. In the following example (from
 9053 Peldszus and Stede, 2013), segment S_2 provides argumentative support for the proposi-
 9054 tion in the segment S_1 :

- 9055 (16.11) [We should tear the building down]_{S1}
9056 [because it is full of asbestos]_{S2}.

9057 Assertions may also support or rebut proposed links between two other assertions, cre-
9058 ating a **hypergraph**, which is a generalization of a graph to the case in which edges can
9059 join any number of vertices. This can be seen by introducing another sentence into the
9060 example:

- 9061 (16.12) [In principle it is possible to clean it up]_{S3}
9062 [but according to the mayor that is too expensive]_{S4}

9063 S3 acknowledges the validity of S2, but **undercuts** its support of S1. This can be repre-
9064 sented by introducing a hyperedge, $(S3, S2, S1)_{\text{undercut}}$, indicating that S3 undercuts the
9065 proposed relationship between S2 and S1. S4 then undercuts the relevance of S3.

9066 **Argumentation mining** is the task of recovering such structures from raw texts. At
9067 present, annotations of argumentation structure are relatively small: Stab and Gurevych
9068 (2014a) have annotated a collection of 90 persuasive essays, and Peldszus and Stede (2015)
9069 have solicited and annotated a set of 112 paragraph-length “microtexts” in German.

9070 16.3.4 Applications of discourse relations

9071 The predominant application of discourse parsing is to select content within a document.
9072 In rhetorical structure theory, the nucleus is considered the more important element of
9073 the relation, and is more likely to be part of a summary of the document; it may also
9074 be more informative for document classification. The D-LTAG theory that underlies the
9075 Penn Discourse Treebank lacks this notion of nuclearity, but arguments may have varying
9076 importance, depending on the relation type. For example, the span of text constituting
9077 ARG1 of an expansion relation is more likely to appear in a summary, while the sentence
9078 constituting ARG2 of an implicit relation is less likely (Louis et al., 2010). Discourse rela-
9079 tions may also signal segmentation points in the document structure. Explicit discourse
9080 markers have been shown to correlate with changes in subjectivity, and identifying such
9081 change points can improve document-level sentiment classification, by helping the clas-
9082 sifier to focus on the subjective parts of the text (Trivedi and Eisenstein, 2013; Yang and
9083 Cardie, 2014).

9084 Extractive Summarization

9085 Text **summarization** is the problem of converting a longer text into a shorter one, while
9086 still conveying the key facts, events, ideas, and sentiments from the original. In **extractive**
9087 **summarization**, the summary is a subset of the original text; in **abstractive summariza-**
9088 **tion**, the summary is produced *de novo*, by paraphrasing the original, or by first encoding

it into a semantic representation (see § 19.2). The main strategy for extractive summarization is to maximize coverage, choosing a subset of the document that best covers the concepts mentioned in the document as a whole; typically, coverage is approximated by bag-of-words overlap (Nenkova and McKeown, 2012). Coverage-based objectives can be supplemented by hierarchical discourse relations, using the principle of nuclearity: in any subordinating discourse relation, the nucleus is more critical to the overall meaning of the text, and is therefore more important to include in an extractive summary (Marcu, 1997a).¹⁰ This insight can be generalized from individual relations using the concept of **discourse depth** (Hirao et al., 2013): for each elementary discourse unit e , the discourse depth d_e is the number of relations in which a discourse unit containing e is the satellite.

Both discourse depth and nuclearity can be incorporated into extractive summarization, using constrained optimization. Let \mathbf{x}_n be a bag-of-words vector representation of elementary discourse unit n , let $y_n \in \{0, 1\}$ indicate whether n is included in the summary, and let d_n be the depth of unit n . Furthermore, let each discourse unit have a “head” h , which is defined recursively:

- if a discourse unit is produced by a subordinating relation, then its head is the head of the (unique) nucleus;
- if a discourse unit is produced by a coordinating relation, then its head is the head of the left-most nucleus;
- for each elementary discourse unit, its parent $\pi(n) \in \{\emptyset, 1, 2, \dots, N\}$ is the head of the smallest discourse unit containing n whose head is not n ;
- if n is the head of the discourse unit spanning the whole document, then $\pi(n) = \emptyset$.

With these definitions in place, discourse-driven extractive summarization can be formalized as (Hirao et al., 2013),

$$\begin{aligned} & \max_{\mathbf{y}=\{0,1\}^N} \sum_{n=1}^N y_n \frac{\Psi(\mathbf{x}_n, \{\mathbf{x}_{1:N}\})}{d_n} \\ & \text{s.t. } \sum_{n=1}^N y_n \left(\sum_{j=1}^V x_{n,j} \right) \leq L \\ & \quad y_{\pi(n)} \geq y_n, \quad \forall n \text{ s.t. } \pi(n) \neq \emptyset \end{aligned} \tag{16.11}$$

where $\Psi(\mathbf{x}_n, \{\mathbf{x}_{1:N}\})$ measures the coverage of elementary discourse unit n with respect to the rest of the document, and $\sum_{j=1}^V x_{n,j}$ is the number of tokens in \mathbf{x}_n . The first constraint ensures that the number of tokens in the summary has an upper bound L . The

¹⁰Conversely, the arguments of a multi-nuclear relation should either both be included in the summary, or both excluded (Durrett et al., 2016).

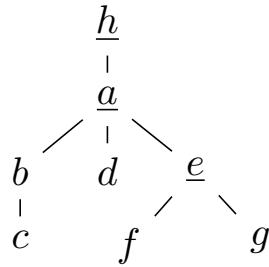


Figure 16.6: A discourse depth tree (Hirao et al., 2013) for the discourse parse from Figure 16.5, in which each elementary discourse unit is connected to its parent. The discourse units in one valid summary are underlined.

second constraint ensures that no elementary discourse unit is included unless its parent is also included. In this way, the discourse structure is used twice: to downweight the contributions of elementary discourse units that are not central to the discourse, and to ensure that the resulting structure is a subtree of the original discourse parse. The optimization problem in 16.11 can be solved with **integer linear programming**, described in § 13.2.2.¹¹

Figure 16.6 shows a discourse depth tree for the RST analysis from Figure 16.5, in which each elementary discourse is connected to (and below) its parent. The underlined discourse units in the figure constitute the following summary:

(16.13) It could have been a great movie, and I really liked the son of the leader of the Samurai. But, other than all that, this movie is nothing more than hidden rip-offs.

Document classification

Hierarchical discourse structures lend themselves naturally to text classification: in a subordinating discourse relation, the nucleus should play a stronger role in the classification decision than the satellite. Various implementations of this idea have been proposed.

- Focusing on within-sentence discourse relations and lexicon-based classification (see § 4.1.2), Voll and Taboada (2007) simply ignore the text in the satellites of each discourse relation.
- At the document level, elements of each discourse relation argument can be reweighted, favoring words in the nucleus, and disfavoring words in the satellite (Heerschap et al., 2011; Bhatia et al., 2015). This approach can be applied recursively, computing

¹¹Formally, 16.11 is a special case of the **knapsack problem**, in which the goal is to find a subset of items with maximum value, constrained by some maximum weight (Cormen et al., 2009).

weights across the entire document. The weights can be relation-specific, so that the features from the satellites of contrastive relations are discounted or even reversed.

- Alternatively, the hierarchical discourse structure can define the structure of a **recursive neural network** (see § 10.6.1). In this network, the representation of each discourse unit is computed from its arguments and from a parameter corresponding to the discourse relation (Ji and Smith, 2017).

Shallow, non-hierarchical discourse relations have also been applied to document classification. One approach is to impose a set of constraints on the analyses of individual discourse units, so that adjacent units have the same polarity when they are connected by a discourse relation indicating agreement, and opposite polarity when connected by a contrastive discourse relation, indicating disagreement (Somasundaran et al., 2009; Zirn et al., 2011). Yang and Cardie (2014) apply explicitly-marked relations from the Penn Discourse Treebank to the problem of sentence-level sentiment polarity classification (see § 4.1). They impose the following soft constraints:

- When a CONTRAST relation appears at the beginning of a sentence, the sentence should have the opposite sentiment polarity as its predecessor.
- When an EXPANSION or CONTINGENCY appears at the beginning of a sentence, it should have the same polarity as its predecessor.
- When a CONTRAST relation appears *within* a sentence, the sentence should have neutral polarity, since it is likely to express both sentiments.

These discourse-driven constraints are shown to improve performance on two datasets of product reviews.

9157 Coherence

Just as **grammaticality** is the property shared by well-structured sentences, **coherence** is the property shared by well-structured discourses. One application of discourse processing is to measure (and maximize) the coherence of computer-generated texts like translations and summaries (Kibble and Power, 2004). Coherence assessment is also used to evaluate human-generated texts, such as student essays (e.g., Miltsakaki and Kukich, 2004; Burstein et al., 2013).

Coherence subsumes a range of phenomena, many of which have been highlighted earlier in this chapter: e.g., that adjacent sentences should be lexically cohesive (Foltz et al., 1998; Ji et al., 2015; Li and Jurafsky, 2017), and that entity references should follow the principles of centering theory (Barzilay and Lapata, 2008; Nguyen and Joty, 2017). Discourse relations also bear on the coherence of a text in a variety of ways:

- Hierarchical discourse relations tend to have a “canonical ordering” of the nucleus and satellite (Mann and Thompson, 1988): for example, in the ELABORATION relation from rhetorical structure theory, the nucleus always comes first, while in the JUSTIFICATION relation, the satellite tends to be first (Marcu, 1997b).
- Discourse relations should be signaled by connectives that are appropriate to the semantic or functional relationship between the arguments: for example, a coherent text would be more likely to use *however* to signal a COMPARISON relation than a *temporal* relation (Kibble and Power, 2004).
- Discourse relations tend to appear in predictable sequences: for example, COMPARISON relations tend to immediately precede CONTINGENCY relations (Pitler et al., 2008). This observation can be formalized by generalizing the entity grid model (§ 16.2.2), so that each cell (i, j) provides information about the role of the discourse argument containing a mention of entity j in sentence i (Lin et al., 2011). For example, if the first sentence is ARG1 of a comparison relation, then any entity mentions in the sentence would be labeled COMP.ARG1. This approach can also be applied to RST discourse relations (Feng et al., 2014).

Datasets One difficulty with evaluating metrics of discourse coherence is that human-generated texts usually meet some minimal threshold of coherence. For this reason, much of the research on measuring coherence has focused on synthetic data. A typical setting is to permute the sentences of a human-written text, and then determine whether the original sentence ordering scores higher according to the proposed coherence measure (Barzilay and Lapata, 2008). There are also small datasets of human evaluations of the coherence of machine summaries: for example, human judgments of the summaries from the participating systems in the 2003 Document Understanding Conference are available online.¹² Researchers from the Educational Testing Service (an organization which administers several national exams in the United States) have studied the relationship between discourse coherence and student essay quality (Burstein et al., 2003, 2010). A public dataset of essays from second-language learners, with quality annotations, has been made available by researchers at Cambridge University (Yannakoudakis et al., 2011). At the other extreme, Louis and Nenkova (2013) analyze the structure of professionally written scientific essays, finding that discourse relation transitions help to distinguish prize-winning essays from other articles in the same genre.

Additional resources

For a manuscript-length discussion of discourse processing, see Stede (2011). Article-length surveys are offered by Webber et al. (2012) and Webber and Joshi (2012).

¹²<http://homepages.inf.ed.ac.uk/mlap/coherence/>

9204 Exercises

- 9205 1. Some discourse connectives tend to occur between their arguments; others can pre-
9206 cede both arguments, and a few can follow both arguments. Indicate whether the
9207 following connectives can occur between, before, and after their arguments: *how-*
9208 *ever, but, while* (contrastive, not temporal), *although, therefore, nonetheless*.
 - 9209 2. This exercise is to be done in pairs. Each participant selects an article from to-
9210 day's news, and replaces all mentions of individual people with special tokens like
9211 PERSON1, PERSON2, and so on. The other participant should then use the rules
9212 of centering theory to guess each type of referring expression: full name (*Captain*
9213 *Ahab*), partial name (e.g., *Ahab*), nominal (e.g., *the ship's captain*), or pronoun. Check
9214 whether the predictions match the original text, and whether the text conforms to
9215 the rules of centering theory.
 - 9216 3. In this exercise, you will produce a figure similar to Figure 16.1.
 - 9217 a) Implement the smoothed cosine similarity metric from Equation 16.2, using the
9218 smoothing kernel $k = [.5, .3, .15, .05]$.
 - 9219 b) Download the text of a news article with at least ten paragraphs.
 - 9220 c) Compute and plot the smoothed similarity \bar{s} over the length of the article.
 - 9221 d) Identify *local minima* in \bar{s} as follows: first find all sentences m such that $\bar{s}_m <$
9222 $\bar{s}_{m \pm 1}$. Then search among these points to find the five sentences with the lowest
9223 \bar{s}_m .
 - 9224 e) How often do the five local minima correspond to paragraph boundaries?
 - 9225 • The fraction of local minima that are paragraph boundaries is the **precision-**
9226 **at- k** , where in this case, $k = 5$.
 - 9227 • The fraction of paragraph boundaries which are local minima is the **recall-**
9228 **at- k** .
 - 9229 • Compute precision-at- k and recall-at- k for $k = 3$ and $k = 10$.
 - 9230 4. One way to formulate text segmentation as a probabilistic model is through the use
9231 of the **Dirichlet Compound Multinomial** (DCM) distribution, which computes the
9232 probability of a bag-of-words, $DCM(\mathbf{x}; \boldsymbol{\alpha})$, where the parameter $\boldsymbol{\alpha}$ is a vector of
9233 positive reals. This distribution can be configured to assign high likelihood to bag-
9234 of-words vectors that are internally coherent, such that individual words appear re-
9235 peatedly: for example, this behavior can be observed for simple parameterizations,
9236 such as $\boldsymbol{\alpha} = \alpha \mathbf{1}$ with $\alpha < 1$.
- 9237 Let $\psi_{\boldsymbol{\alpha}}(i, j)$ represent the log-probability of a segment $w_{i+1:j}$ under a DCM distri-
9238 bution with parameter $\boldsymbol{\alpha}$. Give a dynamic program for segmenting a text into a total

of K segments maximizing the sum of log-probabilities $\sum_{k=1}^K \psi_\alpha(s_{k-1}, s_k)$, where s_k indexes the last token of segment k , and $s_0 = 0$. The time complexity of your dynamic program should not be worse than quadratic in the length of the input and linear in the number of segments.

5. Building on the previous problem, you will now adapt the CKY algorithm to perform hierarchical segmentation. Define a hierarchical segmentation as a set of segmentations $\{\{s_k^{(\ell)}\}_{k=1}^{K^{(\ell)}}\}_{\ell=1}^L$, where L is the segmentation depth. To ensure that the segmentation is hierarchically valid, we require that each segmentation point $s_k^{(\ell)}$ at level ℓ is also a segmentation point at level $\ell - 1$, where $\ell > 1$.

For simplicity, this problem focuses on binary hierarchical segmentation, so that each segment at level $\ell > 1$ has exactly 2 subsegments. Define the score of a hierarchical segmentation as the sum of the scores of all segments (at all levels), using the DCM log-probabilities from the previous problem as the segment scores. Give a CKY-like recurrence such that the optimal “parse” of the text is the maximum log-probability binary segmentation with exactly L levels.

6. The entity grid representation of centering theory can be used to compute a score for adjacent sentences, as described in § 16.2.2. Given a set of sentences, these scores can be used to compute an optimal ordering. Show that finding the ordering with the maximum log probability is NP-complete, by reduction from a well-known problem.
7. In § 16.3.2, it is noted that bottom-up parsing with compositional vector representations of each span is not guaranteed to be optimal. In this exercise, you will construct a minimal example proving this point. Consider a discourse with four units, with base representations $\{z^{(i)}\}_{i=1}^4$. Construct a scenario in which the parse selected by bottom-up parsing is not optimal, and give the precise mathematical conditions under which this suboptimal parse is selected. You may ignore the relation labels ℓ for the purpose of this example.
8. As noted in § 16.3.3, arguments can be described by hypergraphs, in which a segment may **undercut** a proposed edge between two other segments. Extend the model of extractive summarization described in § 16.3.4 to arguments, adding the following constraint: if segment i undercuts an argumentative relationship between j and k , then i cannot be included in the summary unless both j and k are included. Your solution should take the form of a set of *linear* constraints on an integer linear program — that is, each constraint can only involve addition and subtraction of variables.

In the next two exercises, you will explore the use of discourse connectives in a real corpus. Using NLTK, acquire the Brown corpus, and identify sentences that begin with any of the following connectives: *however, nevertheless, moreover, furthermore, thus*.

9276 9. Both lexical consistency and discourse connectives contribute to the **cohesion** of a
9277 text. We might therefore expect adjacent sentences that are joined by explicit dis-
9278 course connectives to also have higher word overlap. Using the Brown corpus, test
9279 this theory by computing the average cosine similarity between adjacent sentences
9280 that are connected by one of the connectives mentioned above. Compare this to the
9281 average cosine similarity of all other adjacent sentences. If you know how, perform
9282 a two-sample t-test to determine whether the observed difference is statistically sig-
9283 nificant.

9284 10. Group the above connectives into the following three discourse relations:

- 9285 • Expansion: *moreover, furthermore*
9286 • Comparison: *however, nevertheless*
9287 • Contingency: *thus*

9288 Focusing on pairs of sentences which are joined by one of these five connectives,
9289 build a classifier to predict the discourse relation from the text of the two adjacent
9290 sentences — taking care to ignore the connective itself. Use the first 30000 sentences
9291 of the Brown corpus as the training set, and the remaining sentences as the test
9292 set. Compare the performance of your classifier against simply choosing the most
9293 common class. Using a bag-of-words classifier, it is hard to do much better than this
9294 baseline, so consider more sophisticated alternatives!

9295

Part IV

9296

Applications

9297

Chapter 17

9298

Information extraction

9299 Computers offer powerful capabilities for searching and reasoning about structured records
9300 and relational data. Some have argued that the most important limitation of artificial in-
9301 telligence is not inference or learning, but simply having too little knowledge (Lenat et al.,
9302 1990). Natural language processing provides an appealing solution: automatically con-
9303 struct a structured **knowledge base** by reading natural language text.

9304 For example, many Wikipedia pages have an “infobox” that provides structured in-
9305 formation about an entity or event. An example is shown in Figure 17.1a: each row rep-
9306 resents one or more properties of the entity IN THE AEROPLANE OVER THE SEA, a record
9307 album. The set of properties is determined by a predefined **schema**, which applies to all
9308 record albums in Wikipedia. As shown in Figure 17.1b, the values for many of these fields
9309 are indicated directly in the first few sentences of text on the same Wikipedia page.

9310 The task of automatically constructing (or “populating”) an infobox from text is an
9311 example of **information extraction**. Much of information extraction can be described in
9312 terms of **entities**, **relations**, and **events**.

9313 • **Entities** are uniquely specified objects in the world, such as people (JEFF MANGUM),
9314 places (ATHENS, GEORGIA), organizations (MERGE RECORDS), and times (FEBRUARY
9315 10, 1998). Chapter 8 described the task of **named entity recognition**, which labels
9316 tokens as parts of entity spans. Now we will see how to go further, **linking** each
9317 entity **mention** to an element in a **knowledge base**.

- 9318 • **Relations** include a **predicate** and two **arguments**: for example, CAPITAL(GEORGIA, ATLANTA).
- **Events** involve multiple typed arguments. For example, the production and release

Studio album by Neutral Milk Hotel	
Released	February 10, 1998
Recorded	July–September 1997
Studio	Pet Sounds Studio, Denver, Colorado
Genre	Indie rock • psychedelic folk • lo-fi
Length	39:55
Label	Merge • Domino
Producer	Robert Schneider

(a) A Wikipedia infobox

- (17.1) In the Aeroplane Over the Sea is the second and final studio album by the American indie rock band Neutral Milk Hotel.
- (17.2) It was released in the United States on February 10, 1998 on Merge Records and May 1998 on Blue Rose Records in the United Kingdom.
- (17.3) Jeff Mangum moved from Athens, Georgia to Denver, Colorado to prepare the bulk of the album's material with producer Robert Schneider, this time at Schneider's newly created Pet Sounds Studio at the home of Jim McIntyre.

- (b) The first few sentences of text. Strings that match fields or field names in the infobox are underlined; strings that mention other entities are wavy underlined.

Figure 17.1: From the Wikipedia page for the album “In the Aeroplane Over the Sea”, retrieved October 26, 2017.

of the album described in Figure 17.1 is described by the event,

```
<TITLE : IN THE AEROPLANE OVER THE SEA,
ARTIST : NEUTRAL MILK HOTEL,
RELEASE-DATE : 1998-FEB-10,...>
```

9319 The set of arguments for an event type is defined by a **schema**. Events often refer to
 9320 time-delimited occurrences: weddings, protests, purchases, terrorist attacks.

9321 Information extraction is similar to semantic role labeling (chapter 13): we may think
 9322 of predicates as corresponding to events, and the arguments as defining slots in the event
 9323 representation. However, the goals of information extraction are different. Rather than
 9324 accurately parsing every sentence, information extraction systems often focus on recog-
 9325 nizing a few key relation or event types, or on the task of identifying all properties of a
 9326 given entity. Information extraction is often evaluated by the correctness of the resulting
 9327 knowledge base, and not by how many sentences were accurately parsed. The goal is
 9328 sometimes described as **macro-reading**, as opposed to **micro-reading**, in which each sen-
 9329 tence must be analyzed correctly. Macro-reading systems are not penalized for ignoring
 9330 difficult sentences, as long as they can recover the same information from other, easier-
 9331 to-read sources. However, macro-reading systems must resolve apparent inconsistencies

9332 (was the album released on MERGE RECORDS or BLUE ROSE RECORDS?), requiring reasoning across the entire dataset.

9334 In addition to the basic tasks of recognizing entities, relations, and events, information
9335 extraction systems must handle negation, and must be able to distinguish statements of
9336 fact from hopes, fears, hunches, and hypotheticals. Finally, information extraction is often paired with the problem of **question answering**, which requires accurately parsing a
9338 query, and then selecting or generating a textual answer. Question answering systems can
9339 be built on knowledge bases that are extracted from large text corpora, or may attempt to
9340 identify answers directly from the source texts.

9341 17.1 Entities

9342 The starting point for information extraction is to identify mentions of entities in text.
9343 Consider the following example:

9344 (17.4) *The United States Army captured a hill overlooking Atlanta on May 14, 1864.*

9345 For this sentence, there are two goals:

- 9346 1. *Identify* the spans *United States Army*, *Atlanta*, and *May 14, 1864* as entity mentions.
9347 (The hill is not uniquely identified, so it is not a *named* entity.) We may also want to
9348 recognize the **named entity types**: organization, location, and date. This is **named**
9349 **entity recognition**, and is described in chapter 8.
- 9350 2. *Link* these spans to entities in a knowledge base: U.S. ARMY, ATLANTA, and 1864-
9351 MAY-14. This task is known as **entity linking**.

9352 The strings to be linked to entities are **mentions** — similar to the use of this term in
9353 coreference resolution. In some formulations of the entity linking task, only named entities
9354 are candidates for linking. This is sometimes called **named entity linking** (Ling et al.,
9355 2015). In other formulations, such as **Wikification** (Milne and Witten, 2008), any string
9356 can be a mention. The set of target entities often corresponds to Wikipedia pages, and
9357 Wikipedia is the basis for more comprehensive knowledge bases such as YAGO (Suchanek
9358 et al., 2007), DBPedia (Auer et al., 2007), and Freebase (Bollacker et al., 2008). Entity link-
9359 ing may also be performed in more “closed” settings, where a much smaller list of targets
9360 is provided in advance. The system must also determine if a mention does not refer to
9361 any entity in the knowledge base, sometimes called a **NIL entity** (McNamee and Dang,
9362 2009).

9363 Returning to (17.4), the three entity mentions may seem unambiguous. But the Wikipedia
9364 disambiguation page for the string *Atlanta* says otherwise:¹ there are more than twenty

¹[https://en.wikipedia.org/wiki/Atlanta_\(disambiguation\)](https://en.wikipedia.org/wiki/Atlanta_(disambiguation)), retrieved November 1, 2017.

9365 different towns and cities, five United States Navy vessels, a magazine, a television show,
 9366 a band, and a singer — each prominent enough to have its own Wikipedia page. We now
 9367 consider how to choose among these dozens of possibilities. In this chapter we will focus
 9368 on supervised approaches. Unsupervised entity linking is closely related to the problem
 9369 of **cross-document coreference resolution**, where the task is to identify pairs of mentions
 9370 that corefer, across document boundaries (Bagga and Baldwin, 1998b; Singh et al., 2011).

9371 17.1.1 Entity linking by learning to rank

9372 Entity linking is often formulated as a **ranking** problem,

$$\hat{y} = \operatorname{argmax}_{y \in \mathcal{Y}(x)} \Psi(y, x, c), \quad [17.1]$$

9373 where y is a target entity, x is a description of the mention, $\mathcal{Y}(x)$ is a set of candidate
 9374 entities, and c is a description of the context — such as the other text in the document,
 9375 or its metadata. The function Ψ is a scoring function, which could be a linear model,
 9376 $\Psi(y, x, c) = \theta \cdot f(y, x, c)$, or a more complex function such as a neural network. In either
 9377 case, the scoring function can be learned by minimizing a margin-based **ranking loss**,

$$\ell(\hat{y}, y^{(i)}, x^{(i)}, c^{(i)}) = (\Psi(\hat{y}, x^{(i)}, c^{(i)}) - \Psi(y^{(i)}, x^{(i)}, c^{(i)}) + 1)_+, \quad [17.2]$$

9378 where $y^{(i)}$ is the ground truth and $\hat{y} \neq y^{(i)}$ is the predicted target for mention $x^{(i)}$ in
 9379 context $c^{(i)}$ (Joachims, 2002; Dredze et al., 2010).

9380 **Candidate identification** For computational tractability, it is helpful to restrict the set of
 9381 candidates, $\mathcal{Y}(x)$. One approach is to use a **name dictionary**, which maps from strings
 9382 to the entities that they might mention. This mapping is many-to-many: a string such as
 9383 *Atlanta* can refer to multiple entities, and conversely, an entity such as ATLANTA can be
 9384 referenced by multiple strings. A name dictionary can be extracted from Wikipedia, with
 9385 links between each Wikipedia entity page and the anchor text of all hyperlinks that point
 9386 to the page (Bunescu and Pasca, 2006; Ratinov et al., 2011). To improve recall, the name
 9387 dictionary can be augmented by partial and approximate matching (Dredze et al., 2010),
 9388 but as the set of candidates grows, the risk of false positives increases. For example, the
 9389 string *Atlanta* is a partial match to *the Atlanta Fed* (a name for the FEDERAL RESERVE BANK
 9390 OF ATLANTA), and a noisy match (edit distance of one) from *Atalanta* (a heroine in Greek
 9391 mythology and an Italian soccer team).

9392 **Features** Feature-based approaches to entity ranking rely on three main types of local
 9393 information (Dredze et al., 2010):

- The similarity of the mention string to the canonical entity name, as quantified by string similarity. This feature would elevate the city ATLANTA over the basketball team ATLANTA HAWKS for the string *Atlanta*.
- The popularity of the entity, which can be measured by Wikipedia page views or PageRank in the Wikipedia link graph. This feature would elevate ATLANTA, GEORGIA over the unincorporated community of ATLANTA, OHIO.
- The entity type, as output by the named entity recognition system. This feature would elevate the city of ATLANTA over the magazine ATLANTA in contexts where the mention is tagged as a location.

In addition to these local features, the document context can also help. If *Jamaica* is mentioned in a document about the Caribbean, it is likely to refer to the island nation; in the context of New York, it is likely to refer to the neighborhood in Queens; in the context of a menu, it might refer to a hibiscus tea beverage. Such hints can be formalized by computing the similarity between the Wikipedia page describing each candidate entity and the mention context $c^{(i)}$, which may include the bag-of-words representing the document (Dredze et al., 2010; Hoffart et al., 2011) or a smaller window of text around the mention (Ratinov et al., 2011). For example, we can compute the cosine similarity between bag-of-words vectors for the context and entity description, typically weighted using **inverse document frequency** to emphasize rare words.²

Neural entity linking An alternative approach is to compute the score for each entity candidate using distributed vector representations of the entities, mentions, and context. For example, for the task of entity linking in Twitter, Yang et al. (2016) employ the bilinear scoring function,

$$\Psi(y, x, c) = v_y^\top \Theta^{(y,x)} x + v_y^\top \Theta^{(y,c)} c, \quad [17.3]$$

with $v_y \in \mathbb{R}^{K_y}$ as the vector embedding of entity y , $x \in \mathbb{R}^{K_x}$ as the embedding of the mention, $c \in \mathbb{R}^{K_c}$ as the embedding of the context, and the matrices $\Theta^{(y,x)}$ and $\Theta^{(y,c)}$ as parameters that score the compatibility of each entity with respect to the mention and context. Each of the vector embeddings can be learned from an end-to-end objective, or pre-trained on unlabeled data.

- Pretrained **entity embeddings** can be obtained from an existing knowledge base (Bordes et al., 2011, 2013), or by running a word embedding algorithm such as WORD2VEC

²The **document frequency** of word j is $DF(j) = \frac{1}{N} \sum_{i=1}^N \delta(x_j^{(i)} > 0)$, equal to the number of documents in which the word appears. The contribution of each word to the cosine similarity of two bag-of-words vectors can be weighted by the **inverse document frequency** $\frac{1}{DF(j)}$ or $\log \frac{1}{DF(j)}$, to emphasize rare words (Spärck Jones, 1972).

- 9424 on the text of Wikipedia, with hyperlinks substituted for the anchor text.³
- 9425 • The embedding of the mention x can be computed by averaging the embeddings
 9426 of the words in the mention (Yang et al., 2016), or by the compositional techniques
 9427 described in § 14.8.
- 9428 • The embedding of the context c can also be computed from the embeddings of the
 9429 words in the context. A **denoising autoencoder** learns a function from raw text to
 9430 dense K -dimensional vector encodings by minimizing a reconstruction loss (Vin-
 9431 cent et al., 2010),

$$\min_{\theta_g, \theta_h} \sum_{i=1}^N \|\mathbf{x}^{(i)} - g(h(\tilde{\mathbf{x}}^{(i)}; \theta_h); \theta_g)\|^2, \quad [17.4]$$

9432 where $\tilde{\mathbf{x}}^{(i)}$ is a noisy version of the bag-of-words counts $\mathbf{x}^{(i)}$, which is produced by
 9433 randomly setting some counts to zero; $h : \mathbb{R}^V \rightarrow \mathbb{R}^K$ is an encoder with parameters
 9434 θ_h ; and $g : \mathbb{R}^K \rightarrow \mathbb{R}^V$, with parameters θ_g . The encoder and decoder functions
 9435 are typically implemented as feedforward neural networks. To apply this model to
 9436 entity linking, each entity and context are initially represented by the encoding of
 9437 their bag-of-words vectors, $h(e)$ and $g(c)$, and these encodings are then fine-tuned
 9438 from labeled data (He et al., 2013). The context vector c can also be obtained by
 9439 convolution (§ 3.4) on the embeddings of words in the document (Sun et al., 2015),
 9440 or by examining metadata such as the author’s social network (Yang et al., 2016).

9441 The remaining parameters $\Theta^{(y,x)}$ and $\Theta^{(y,c)}$ can be trained by backpropagation from the
 9442 margin loss in Equation 17.2.

9443 17.1.2 Collective entity linking

9444 Entity linking can be more accurate when it is performed jointly across a document. To
 9445 see why, consider the following lists:

- 9446 (17.5) a. California, Oregon, Washington
 9447 b. Baltimore, Washington, Philadelphia
 9448 c. Washington, Adams, Jefferson

9449 In each case, the term *Washington* refers to a different entity, and this reference is strongly
 9450 suggested by the other entries on the list. In the last list, all three names are highly am-
 9451 biguous — there are dozens of other *Adams* and *Jefferson* entities in Wikipedia. But a

³Pre-trained entity embeddings can be downloaded from <https://code.google.com/archive/p/word2vec/>.

9448 preference for coherence motivates **collectively** linking these references to the first three
 9449 U.S. presidents.

9450 A general approach to collective entity linking is to introduce a compatibility score
 9451 $\psi_c(\mathbf{y})$. Collective entity linking is then performed by optimizing the global objective,

$$\hat{\mathbf{y}} = \underset{\mathbf{y} \in \mathbb{Y}(\mathbf{x})}{\operatorname{argmax}} \Psi_c(\mathbf{y}) + \sum_{i=1}^N \Psi_\ell(y^{(i)}, \mathbf{x}^{(i)}, \mathbf{c}^{(i)}), \quad [17.5]$$

9452 where $\mathbb{Y}(\mathbf{x})$ is the set of all possible collective entity assignments for the mentions in \mathbf{x} ,
 9453 and ψ_ℓ is the local scoring function for each entity i . The compatibility function is typically
 9454 decomposed into a sum of pairwise scores, $\Psi_c(\mathbf{y}) = \sum_{i=1}^N \sum_{j \neq i}^N \Psi_c(y^{(i)}, y^{(j)})$. These scores
 9455 can be computed in a number of different ways:

- 9456 • Wikipedia defines high-level categories for entities (e.g., *living people*, *Presidents of*
 9457 *the United States*, *States of the United States*), and Ψ_c can reward entity pairs for the
 9458 number of categories that they have in common (Cucerzan, 2007).
- 9459 • Compatibility can be measured by the number of incoming hyperlinks shared by
 9460 the Wikipedia pages for the two entities (Milne and Witten, 2008).
- 9461 • In a neural architecture, the compatibility of two entities can be set equal to the inner
 9462 product of their embeddings, $\Psi_c(y^{(i)}, y^{(j)}) = \mathbf{v}_{y^{(i)}} \cdot \mathbf{v}_{y^{(j)}}$.
- 9463 • A non-pairwise compatibility score can be defined using a type of latent variable
 9464 model known as a **probabilistic topic model** (Blei et al., 2003; Blei, 2012). In this
 9465 framework, each latent topic is a probability distribution over entities, and each
 9466 document has a probability distribution over topics. Each entity helps to determine
 9467 the document's distribution over topics, and in turn these topics help to resolve am-
 9468 biguous entity mentions (Newman et al., 2006). Inference can be performed using
 9469 the sampling techniques described in chapter 5.

9470 Unfortunately, collective entity linking is **NP-hard** even for pairwise compatibility func-
 9471 tions, so exact optimization is almost certainly intractable. Various approximate inference
 9472 techniques have been proposed, including **integer linear programming** (Cheng and Roth,
 9473 2013), **Gibbs sampling** (Han and Sun, 2012), and graph-based algorithms (Hoffart et al.,
 9474 2011; Han et al., 2011).

9475 17.1.3 *Pairwise ranking loss functions

9476 The loss function defined in Equation 17.2 considers only the highest-scoring prediction
 9477 \hat{y} , but in fact, the true entity $y^{(i)}$ should outscore *all* other entities. A loss function based on
 9478 this idea would give a gradient against the features or representations of several entities,

Algorithm 18 WARP approximate ranking loss

```

1: procedure WARP( $y^{(i)}$ ,  $\mathbf{x}^{(i)}$ )
2:    $N \leftarrow 0$ 
3:   repeat
4:     Randomly sample  $y \sim \mathcal{Y}(\mathbf{x}^{(i)})$ 
5:      $N \leftarrow N + 1$ 
6:     if  $\psi(y, \mathbf{x}^{(i)}) + 1 > \psi(y^{(i)}, \mathbf{x}^{(i)})$  then            $\triangleright$  check for margin violation
7:        $r \leftarrow \lfloor |\mathcal{Y}(\mathbf{x}^{(i)})|/N \rfloor$                           $\triangleright$  compute approximate rank
8:       return  $L_{\text{rank}}(r) \times (\psi(y, \mathbf{x}^{(i)}) + 1 - \psi(y^{(i)}, \mathbf{x}^{(i)}))$ 
9:     until  $N \geq |\mathcal{Y}(\mathbf{x}^{(i)})| - 1$                             $\triangleright$  no violation found
10:    return 0                                          $\triangleright$  return zero loss

```

9479 not just the top-scoring prediction. Usunier et al. (2009) define a general ranking error
 9480 function,

$$L_{\text{rank}}(k) = \sum_{j=1}^k \alpha_j, \quad \text{with } \alpha_1 \geq \alpha_2 \geq \dots \geq 0, \quad [17.6]$$

9481 where k is equal to the number of labels ranked higher than the correct label $y^{(i)}$. This
 9482 function defines a class of ranking errors: if $\alpha_j = 1$ for all j , then the ranking error is
 9483 equal to the rank of the correct entity; if $\alpha_1 = 1$ and $\alpha_{j>1} = 0$, then the ranking error is
 9484 one whenever the correct entity is not ranked first; if α_j decreases smoothly with j , as in
 9485 $\alpha_j = \frac{1}{j}$, then the error is between these two extremes.

This ranking error can be integrated into a margin objective. Remember that large margin classification requires not only the correct label, but also that the correct label outscores other labels by a substantial margin. A similar principle applies to ranking: we want a high rank for the correct entity, and we want it to be separated from other entities by a substantial margin. We therefore define the margin-augmented rank,

$$r(y^{(i)}, \mathbf{x}^{(i)}) \triangleq \sum_{y \in \mathcal{Y}(\mathbf{x}^{(i)}) \setminus y^{(i)}} \delta \left(1 + \psi(y, \mathbf{x}^{(i)}) \geq \psi(y^{(i)}, \mathbf{x}^{(i)}) \right), \quad [17.7]$$

9486 where $\delta(\cdot)$ is a delta function, and $\mathcal{Y}(\mathbf{x}^{(i)}) \setminus y^{(i)}$ is the set of all entity candidates minus
 9487 the true entity $y^{(i)}$. The margin-augmented rank is the rank of the true entity, after aug-
 9488 menting every other candidate with a margin of one, under the current scoring function
 9489 ψ . (The context c is omitted for clarity, and can be considered part of x .)

For each instance, a hinge loss is computed from the ranking error associated with this

margin-augmented rank, and the violation of the margin constraint,

$$\ell(y^{(i)}, \mathbf{x}^{(i)}) = \frac{L_{\text{rank}}(r(y^{(i)}, \mathbf{x}^{(i)}))}{r(y^{(i)}, \mathbf{x}^{(i)})} \sum_{y \in \mathcal{Y}(\mathbf{x}) \setminus y^{(i)}} \left(\psi(y, \mathbf{x}^{(i)}) - \psi(y^{(i)}, \mathbf{x}^{(i)}) + 1 \right)_+, \quad [17.8]$$

The sum in Equation 17.8 includes non-zero values for every label that is ranked at least as high as the true entity, after applying the margin augmentation. Dividing by the margin-augmented rank of the true entity thus gives the average violation.

The objective in Equation 17.8 is expensive to optimize when the label space is large, as is usually the case for entity linking against large knowledge bases. This motivates a randomized approximation called **WARP** (Weston et al., 2011), shown in Algorithm 18. In this procedure, we sample random entities until one violates the pairwise margin constraint, $\psi(y, \mathbf{x}^{(i)}) + 1 \geq \psi(y^{(i)}, \mathbf{x}^{(i)})$. The number of samples N required to find such a violation yields an approximation of the margin-augmented rank of the true entity, $r(y^{(i)}, \mathbf{x}^{(i)}) \approx \left\lceil \frac{|\mathcal{Y}(\mathbf{x})|}{N} \right\rceil$. If a violation is found immediately, $N = 1$, the correct entity probably ranks below many others, $r \approx |\mathcal{Y}(\mathbf{x})|$. If many samples are required before a violation is found, $N \rightarrow |\mathcal{Y}(\mathbf{x})|$, then the correct entity is probably highly ranked, $r \rightarrow 1$. A computational advantage of WARP is that it is not necessary to find the highest-scoring label, which can impose a non-trivial computational cost when $\mathcal{Y}(\mathbf{x}^{(i)})$ is large. The objective is conceptually similar to the **negative sampling** objective in WORD2VEC (chapter 14), which compares the observed word against randomly sampled alternatives.

17.2 Relations

After identifying the entities that are mentioned in a text, the next step is to determine how they are related. Consider the following example:

(17.6) George Bush traveled to France on Thursday for a summit.

This sentence introduces a relation between the entities referenced by *George Bush* and *France*. In the Automatic Content Extraction (ACE) ontology (Linguistic Data Consortium, 2005), the type of this relation is PHYSICAL, and the subtype is LOCATED. This relation would be written,

PHYSICAL.LOCATED(GEORGE BUSH, FRANCE). [17.9]

Relations take exactly two arguments, and the order of the arguments matters.

In the ACE datasets, relations are annotated between entity mentions, as in the example above. Relations can also hold between nominals, as in the following example from the SemEval-2010 shared task (Hendrickx et al., 2009):

CAUSE-EFFECT	<i>those cancers were caused by radiation exposures</i>
INSTRUMENT-AGENCY	<i>phone operator</i>
PRODUCT-PRODUCER	<i>a factory manufactures suits</i>
CONTENT-CONTAINER	<i>a bottle of honey was weighed</i>
ENTITY-ORIGIN	<i>letters from foreign countries</i>
ENTITY-DESTINATION	<i>the boy went to bed</i>
COMPONENT-WHOLE	<i>my apartment has a large kitchen</i>
MEMBER-COLLECTION	<i>there are many trees in the forest</i>
COMMUNICATION-TOPIC	<i>the lecture was about semantics</i>

Table 17.1: Relations and example sentences from the SemEval-2010 dataset (Hendrickx et al., 2009)

9518 (17.7) The cup contained tea from dried ginseng.

9519 This sentence describes a relation of type ENTITY-ORIGIN between *tea* and *ginseng*. Nominal
 9520 relation extraction is closely related to **semantic role labeling** (chapter 13). The main
 9521 difference is that relation extraction is restricted to a relatively small number of relation
 9522 types; for example, Table 17.1 shows the ten relation types from SemEval-2010.

9523 17.2.1 Pattern-based relation extraction

9524 Early work on relation extraction focused on hand-crafted patterns (Hearst, 1992). For
 9525 example, the appositive *Starbuck, a native of Nantucket* signals the relation ENTITY-ORIGIN
 9526 between *Starbuck* and *Nantucket*. This pattern can be written as,

$$\text{PERSON , } a \text{ native of LOCATION} \Rightarrow \text{ENTITY-ORIGIN(PERSON, LOCATION)}. \quad [17.10]$$

9527 This pattern will be “triggered” whenever the literal string *, a native of* occurs between an
 9528 entity of type PERSON and an entity of type LOCATION. Such patterns can be generalized
 9529 beyond literal matches using techniques such as lemmatization, which would enable the
 9530 words (*buy, buys, buying*) to trigger the same patterns (see § 4.3.1). A more aggressive
 9531 strategy would be to group all words in a WordNet synset (§ 4.2), so that, e.g., *buy* and
 9532 *purchase* trigger the same patterns.

9533 Relation extraction patterns can be implemented in finite-state automata (§ 9.1). If the
 9534 named entity recognizer is also a finite-state machine, then the systems can be combined
 9535 by finite-state transduction (Hobbs et al., 1997). This makes it possible to propagate uncer-
 9536 tainty through the finite-state cascade, and disambiguate from higher-level context. For
 9537 example, suppose the entity recognizer cannot decide whether *Starbuck* refers to either a
 9538 PERSON or a LOCATION; in the composed transducer, the relation extractor would be free
 9539 to select the PERSON annotation when it appears in the context of an appropriate pattern.

9540 **17.2.2 Relation extraction as a classification task**

9541 Relation extraction can be formulated as a classification problem,

$$\hat{r}_{(i,j),(m,n)} = \operatorname{argmax}_{r \in \mathcal{R}} \Psi(r, (i, j), (m, n), \mathbf{w}), \quad [17.11]$$

9542 where $r \in \mathcal{R}$ is a relation type (possibly NIL), $\mathbf{w}_{i+1:j}$ is the span of the first argument, and
 9543 $\mathbf{w}_{m+1:n}$ is the span of the second argument. The argument $\mathbf{w}_{m+1:n}$ may appear before
 9544 or after $\mathbf{w}_{i+1:j}$ in the text, or they may overlap; we stipulate only that $\mathbf{w}_{i+1:j}$ is the first
 9545 argument of the relation. We now consider three alternatives for computing the scoring
 9546 function.

9547 **Feature-based classification**

9548 In a feature-based classifier, the scoring function is defined as,

$$\Psi(r, (i, j), (m, n), \mathbf{w}) = \boldsymbol{\theta} \cdot \mathbf{f}(r, (i, j), (m, n), \mathbf{w}), \quad [17.12]$$

9549 with $\boldsymbol{\theta}$ representing a vector of weights, and $\mathbf{f}(\cdot)$ a vector of features. The pattern-based
 9550 methods described in § 17.2.1 suggest several features:

- 9551 • Local features of $\mathbf{w}_{i+1:j}$ and $\mathbf{w}_{m+1:n}$, including: the strings themselves; whether they
 9552 are recognized as entities, and if so, which type; whether the strings are present in a
 9553 **gazetteer** of entity names; each string's syntactic head (§ 9.2.2).
- 9554 • Features of the span between the two arguments, $\mathbf{w}_{j+1:m}$ or $\mathbf{w}_{n+1:i}$ (depending on
 9555 which argument appears first): the length of the span; the specific words that appear
 9556 in the span, either as a literal sequence or a bag-of-words; the wordnet synsets (§ 4.2)
 9557 that appear in the span between the arguments.
- 9558 • Features of the syntactic relationship between the two arguments, typically the **de-**
 9559 **pendency path** between the arguments (§ 13.2.1). Example dependency paths are
 9560 shown in Table 17.2.

9561 **Kernels**

9562 Suppose that the first line of Table 17.2 is a labeled example, and the remaining lines are
 9563 instances to be classified. A feature-based approach would have to decompose the depen-
 9564 dency paths into features that capture individual edges, with or without their labels, and
 9565 then learn weights for each of these features: for example, the second line contains identi-
 9566 cal dependencies, but different arguments; the third line contains a different inflection of
 9567 the word *travel*; the fourth and fifth lines each contain an additional edge on the depen-
 9568 dency path; and the sixth example uses an entirely different path. Rather than attempting
 9569 to create local features that capture all of the ways in which these dependencies paths

1. <i>George Bush traveled to France</i>	<i>George Bush</i> \leftarrow traveled \rightarrow France NSUBJ OBL
2. <i>Ahab traveled to Nantucket</i>	<i>Ahab</i> \leftarrow traveled \rightarrow Nantucket NSUBJ OBL
3. <i>George Bush will travel to France</i>	<i>George Bush</i> \leftarrow travel \rightarrow France NSUBJ OBL
4. <i>George Bush wants to travel to France</i>	<i>George Bush</i> \leftarrow wants \rightarrow travel \rightarrow France NSUBJ XCOMP OBL
5. <i>Ahab traveled to a city in France</i>	<i>Ahab</i> \leftarrow traveled \rightarrow city \rightarrow France NSUBJ OBL NMOD
6. <i>We await Ahab's visit to France</i>	<i>Ahab</i> \leftarrow visit \rightarrow France NMOD:POSS NMOD

Table 17.2: Candidates instances for the PHYSICAL.LOCATED relation, and their dependency paths

9570 are similar and different, we can instead define a similarity function κ , which computes a
 9571 score for any pair of instances, $\kappa : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_+$. The score for any pair of instances (i, j)
 9572 is $\kappa(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) \geq 0$, with $\kappa(i, j)$ being large when instances $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(j)}$ are similar. If the
 9573 function κ obeys a few key properties it is a valid **kernel function**.⁴

Given a valid kernel function, we can build a non-linear classifier without explicitly defining a feature vector or neural network architecture. For a binary classification problem $y \in \{-1, 1\}$, we have the decision function,

$$\hat{y} = \text{Sign}(b + \sum_{i=1}^N y^{(i)} \alpha^{(i)} \kappa(\mathbf{x}^{(i)}, \mathbf{x})) \quad [17.13]$$

9574 where b and $\{\alpha^{(i)}\}_{i=1}^N$ are parameters that must be learned from the training set, under
 9575 the constraint $\forall_i, \alpha^{(i)} \geq 0$. Intuitively, each α_i specifies the importance of the instance $\mathbf{x}^{(i)}$
 9576 towards the classification rule. Kernel-based classification can be viewed as a weighted
 9577 form of the **nearest-neighbor** classifier (Hastie et al., 2009), in which test instances are
 9578 assigned the most common label among their near neighbors in the training set. This
 9579 results in a non-linear classification boundary. The parameters are typically learned from
 9580 a margin-based objective (see § 2.4), leading to the **kernel support vector machine**. To
 9581 generalize to multi-class classification, we can train separate binary classifiers for each
 9582 label (sometimes called **one-versus-all**), or train binary classifiers for each pair of possible
 9583 labels (**one-versus-one**).

9584 Dependency kernels are particularly effective for relation extraction, due to their ability
 9585 to capture syntactic properties of the path between the two candidate arguments. One
 9586 class of dependency tree kernels is defined recursively, with the score for a pair of trees

⁴The **Gram matrix** \mathbf{K} arises from computing the kernel function between all pairs in a set of instances. For a valid kernel, the Gram matrix must be symmetric ($\mathbf{K} = \mathbf{K}^\top$) and positive semi-definite ($\forall \mathbf{a}, \mathbf{a}^\top \mathbf{K} \mathbf{a} \geq 0$). For more on kernel-based classification, see chapter 14 of Murphy (2012).

equal to the similarity of the root nodes and the sum of similarities of matched pairs of child subtrees (Zelenko et al., 2003; Culotta and Sorensen, 2004). Alternatively, Bunescu and Mooney (2005) define a kernel function over sequences of unlabeled dependency edges, in which the score is computed as a product of scores for each pair of words in the sequence: identical words receive a high score, words that share a synset or part-of-speech receive a small non-zero score (e.g., *travel* / *visit*), and unrelated words receive a score of zero.

Neural relation extraction

Convolutional neural networks (§ 3.4) were an early neural architecture for relation extraction (Zeng et al., 2014; dos Santos et al., 2015). For the sentence (w_1, w_2, \dots, w_M) , obtain a matrix of word embeddings \mathbf{X} , where $\mathbf{x}_m \in \mathbb{R}^K$ is the embedding of w_m . Now, suppose the candidate arguments appear at positions a_1 and a_2 ; then for each word in the sentence, its position with respect to each argument is $m - a_1$ and $m - a_2$. (Following Zeng et al. (2014), this is a restricted version of the relation extraction task in which the arguments are single tokens.) To capture any information conveyed by these positions, the word embeddings are concatenated with vector encodings of the positional offsets, $\mathbf{x}_{m-a_1}^{(p)}$ and $\mathbf{x}_{m-a_2}^{(p)}$. (For more on **positional encodings**, see § 18.3.2.) The complete base representation of the sentence is,

$$\mathbf{X}(a_1, a_2) = \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_M \\ \mathbf{x}_{1-a_1}^{(p)} & \mathbf{x}_{2-a_1}^{(p)} & \cdots & \mathbf{x}_{M-a_1}^{(p)} \\ \mathbf{x}_{1-a_2}^{(p)} & \mathbf{x}_{2-a_2}^{(p)} & \cdots & \mathbf{x}_{M-a_2}^{(p)} \end{pmatrix}, \quad [17.14]$$

where each column is a vertical concatenation of a word embedding, represented by the column vector \mathbf{x}_m , and two positional encodings, specifying the position with respect to a_1 and a_2 . The matrix $\mathbf{X}(a_1, a_2)$ is then taken as input to a convolutional layer (see § 3.4), and max-pooling is applied to obtain a vector. The final scoring function is then,

$$\Psi(r, i, j, \mathbf{X}) = \theta_r \cdot \text{MaxPool}(\text{ConvNet}(\mathbf{X}(i, j); \phi)), \quad [17.15]$$

where ϕ defines the parameters of the convolutional operator, and the θ_r defines a set of weights for relation r . The model can be trained using a margin objective,

$$\hat{r} = \underset{r}{\operatorname{argmax}} \Psi(r, i, j, \mathbf{X}) \quad [17.16]$$

$$\ell = (1 + \psi(\hat{r}, i, j, \mathbf{X}) - \psi(r, i, j, \mathbf{X}))_+. \quad [17.17]$$

Recurrent neural networks (§ 6.3) have also been applied to relation extraction, using a network such as a bidirectional LSTM to encode the words or dependency path between the two arguments. Xu et al. (2015) segment each dependency path into left and

right subpaths: the path $George \underset{\text{NSUBJ}}{Bush} \leftarrow \underset{\text{XCOMP}}{wants} \rightarrow \underset{\text{OBL}}{travel} \rightarrow France$ is segmented into the subpaths, $George \underset{\text{NSUBJ}}{Bush} \leftarrow \underset{\text{XCOMP}}{wants}$ and $wants \rightarrow \underset{\text{XCOMP}}{travel} \rightarrow \underset{\text{OBL}}{France}$. In each path, a recurrent neural network is run from the argument to the root word (in this case, *wants*). The final representation by max pooling (§ 3.4) across all the recurrent states along each path. This process can be applied across separate “channels”, in which the inputs consist of embeddings for the words, parts-of-speech, dependency relations, and WordNet hypernyms (e.g., *France-nation*; see § 4.2). To define the model formally, let $s(m)$ define the successor of word m in either the left or right subpath (in a dependency path, each word can have a successor in at most one subpath). Let $\mathbf{x}_m^{(c)}$ indicate the embedding of word (or relation) m in channel c , and let $\overleftarrow{\mathbf{h}}_m^{(c)}$ and $\overrightarrow{\mathbf{h}}_m^{(c)}$ indicate the associated recurrent states in the left and right subtrees respectively. Then the complete model is specified as follows,

$$\mathbf{h}_{s(m)}^{(c)} = \text{RNN}(\mathbf{x}_{s(m)}^{(c)}, \mathbf{h}_m^{(c)}) \quad [17.18]$$

$$\mathbf{z}^{(c)} = \text{MaxPool}(\overleftarrow{\mathbf{h}}_i^{(c)}, \overleftarrow{\mathbf{h}}_{s(i)}^{(c)}, \dots, \overleftarrow{\mathbf{h}}_{\text{root}}^{(c)}, \overrightarrow{\mathbf{h}}_j^{(c)}, \overrightarrow{\mathbf{h}}_{s(j)}^{(c)}, \dots, \overrightarrow{\mathbf{h}}_{\text{root}}^{(c)}) \quad [17.19]$$

$$\Psi(r, i, j) = \boldsymbol{\theta} \cdot [\mathbf{z}^{(\text{word})}; \mathbf{z}^{(\text{POS})}; \mathbf{z}^{(\text{dependency})}; \mathbf{z}^{(\text{hypernym})}] \quad [17.20]$$

9609 Note that \mathbf{z} is computed by applying max-pooling to the *matrix* of horizontally concatenated vectors \mathbf{h} , while Ψ is computed from the *vector* of vertically concatenated vectors
9610 \mathbf{z} . Xu et al. (2015) pass the score Ψ through a **softmax** layer to obtain a probability
9611 $p(r | i, j, \mathbf{w})$, and train the model by regularized **cross-entropy**. Miwa and Bansal (2016)
9612 show that a related model can solve the more challenging “end-to-end” relation extrac-
9613 tion task, in which the model must simultaneously detect entities and then extract their
9614 relations.

9616 17.2.3 Knowledge base population

9617 In many applications, what matters is not what fraction of sentences are analyzed cor-
9618 rectly, but how much accurate knowledge can be extracted. **Knowledge base population**
9619 (**KBP**) refers to the task of filling in Wikipedia-style infoboxes, as shown in Figure 17.1a.
9620 Knowledge base population can be decomposed into two subtasks: **entity linking** (de-
9621 scribed in § 17.1), and **slot filling** (Ji and Grishman, 2011). Slot filling has two key dif-
9622 ferences from the formulation of relation extraction presented above: the relations hold
9623 between entities rather than spans of text, and the performance is evaluated at the *type*
9624 *level* (on entity pairs), rather than on the *token level* (on individual sentences).

9625 From a practical standpoint, there are three other important differences between slot
9626 filling and per-sentence relation extraction.

- 9627 • KBP tasks are often formulated from the perspective of identifying attributes of a
9628 few “query” entities. As a result, these systems often start with an **information**

- 9629 **retrieval** phase, in which relevant passages of text are obtained by search.
9630 • For many entity pairs, there will be multiple passages of text that provide evidence.
9631 Slot filling systems must aggregate this evidence to predict a single relation type (or
9632 set of relations).
9633 • Labeled data is usually available in the form of pairs of related entities, rather than
9634 annotated passages of text. Training from such type-level annotations is a challenge:
9635 two entities may be linked by several relations, or they may appear together in a
9636 passage of text that nonetheless does not describe their relation to each other.

9637 Information retrieval is beyond the scope of this text (see Manning et al., 2008). The re-
9638 mainder of this section describes approaches to information fusion and learning from
9639 type-level annotations.

9640 **Information fusion**

9641 In knowledge base population, there will often be multiple pieces of evidence for (and
9642 sometimes against) a single relation. For example, a search for the entity MAYNARD JACK-
9643 SON, JR. may return several passages that reference the entity ATLANTA:⁵

- 9644 (17.8) a. Elected mayor of **Atlanta** in 1973, **Maynard Jackson** was the first African
9645 American to serve as mayor of a major southern city.
9646 b. **Atlanta**'s airport will be renamed to honor **Maynard Jackson**, the city's first
9647 Black mayor.
9648 c. Born in Dallas, Texas in 1938, **Maynard Holbrook Jackson, Jr.** moved to
9649 **Atlanta** when he was 8.
9650 d. **Maynard Jackson** has gone from one of the worst high schools in **Atlanta** to
9651 one of the best.

9652 The first and second examples provide evidence for the relation **MAYOR** holding between
9653 the entities **ATLANTA** and **MAYNARD JACKSON, JR.**. The third example provides evidence
9654 for a different relation between these same entities, **LIVED-IN**. The fourth example poses
9655 an entity linking problem, referring to **MAYNARD JACKSON HIGH SCHOOL**. Knowledge
9656 base population requires aggregating this sort of textual evidence, and predicting the re-
9657 lations that are most likely to hold.

9658 One approach is to run a single-document relation extraction system (using the tech-
9659 niques described in § 17.2.2), and then aggregate the results (Li et al., 2011). Relations

⁵First three examples from: <http://www.georgiaencyclopedia.org/articles/government-politics/maynard-jackson-1938-2003>; JET magazine, November 10, 2003; www.todayingeorgiahistory.org/content/maynard-jackson-elected

9660 that are detected with high confidence in multiple documents are more likely to be valid,
 9661 motivating the heuristic,

$$\psi(r, e_1, e_2) = \sum_{i=1}^N (\text{p}(r(e_1, e_2) | \mathbf{w}^{(i)}))^{\alpha}, \quad [17.21]$$

9662 where $\text{p}(r(e_1, e_2) | \mathbf{w}^{(i)})$ is the probability of relation r between entities e_1 and e_2 conditioned
 9663 on the text $\mathbf{w}^{(i)}$, and $\alpha \gg 1$ is a tunable hyperparameter. Using this heuristic, it is
 9664 possible to rank all candidate relations, and trace out a **precision-recall curve** as more re-
 9665 lations are extracted.⁶ Alternatively, features can be aggregated across multiple passages
 9666 of text, feeding a single type-level relation extraction system (Wolfe et al., 2017).

9667 Precision can be improved by introducing constraints across multiple relations. For
 9668 example, if we are certain of the relation $\text{PARENT}(e_1, e_2)$, then it cannot also be the case
 9669 that $\text{PARENT}(e_2, e_1)$. Integer linear programming makes it possible to incorporate such
 9670 constraints into a global optimization (Li et al., 2011). Other pairs of relations have pos-
 9671 itive correlations, such $\text{MAYOR}(e_1, e_2)$ and $\text{LIVED-IN}(e_1, e_2)$. Compatibility across relation
 9672 types can be incorporated into probabilistic graphical models (e.g., Riedel et al., 2010).

9673 Distant supervision

9674 Relation extraction is “annotation hungry,” because each relation requires its own la-
 9675 beled data. Rather than relying on annotations of individual documents, it would be
 9676 preferable to use existing knowledge resources — such as the many facts that are al-
 9677 ready captured in knowledge bases like DBpedia. However such annotations raise the
 9678 inverse of the information fusion problem considered above: the existence of the relation
 9679 $\text{MAYOR}(\text{MAYNARD JACKSON JR., ATLANTA})$ provides only **distant supervision** for the
 9680 example texts in which this entity pair is mentioned.

9681 One approach is to treat the entity pair as the instance, rather than the text itself (Mintz
 9682 et al., 2009). Features are then aggregated across all sentences in which both entities are
 9683 mentioned, and labels correspond to the relation (if any) between the entities in a knowl-
 9684 edge base, such as FreeBase. Negative instances are constructed from entity pairs that are
 9685 not related in the knowledge base. In some cases, two entities are related, but the knowl-
 9686 edge base is missing the relation; however, because the number of possible entity pairs is
 9687 huge, these missing relations are presumed to be relatively rare. This approach is shown
 9688 in Figure 17.2.

9689 In **multiple instance learning**, labels are assigned to *sets* of instances, of which only
 9690 an unknown subset are actually relevant (Dietterich et al., 1997; Maron and Lozano-Pérez,
 9691 1998). This formalizes the framework of distant supervision: the relation $\text{REL}(A, B)$ acts

⁶The precision-recall curve is similar to the ROC curve shown in Figure 4.4, but it includes the precision $\frac{\text{TP}}{\text{TP} + \text{FP}}$ rather than the false positive rate $\frac{\text{FP}}{\text{FP} + \text{TN}}$.

- **Label** : MAYOR(ATLANTA, MAYNARD JACKSON)
 - Elected mayor of **Atlanta** in 1973, **Maynard Jackson** ...
 - **Atlanta**'s airport will be renamed to honor **Maynard Jackson**, the city's first Black mayor
 - Born in Dallas, Texas in 1938, **Maynard Holbrook Jackson, Jr.** moved to **Atlanta** when he was 8.
- **Label** : MAYOR(NEW YORK, FIORELLO LA GUARDIA)
 - **Fiorello La Guardia** was Mayor of **New York** for three terms ...
 - **Fiorello La Guardia**, then serving on the **New York** City Board of Aldermen...
- **Label** : BORN-IN(DALLAS, MAYNARD JACKSON)
 - Born in **Dallas**, Texas in 1938, **Maynard Holbrook Jackson, Jr.** moved to Atlanta when he was 8.
 - **Maynard Jackson** was raised in **Dallas** ...
- **Label** : NIL(NEW YORK, MAYNARD JACKSON)
 - **Jackson** married Valerie Richardson, whom he had met in **New York**...
 - **Jackson** was a member of the Georgia and **New York** bars ...

Figure 17.2: Four training instances for relation classification using **distant supervision** Mintz et al. (2009). The first two instances are positive for the MAYOR relation, and the third instance is positive for the BORN-IN relation. The fourth instance is a negative example, constructed from a pair of entities (NEW YORK, MAYNARD JACKSON) that do not appear in any Freebase relation. Each instance's features are computed by aggregating across all sentences in which the two entities are mentioned.

9692 as a label for the entire set of sentences mentioning entities A and B, even when only a
 9693 subset of these sentences actually describes the relation. One approach to multi-instance
 9694 learning is to introduce a binary **latent variable** for each sentence, indicating whether the
 9695 sentence expresses the labeled relation (Riedel et al., 2010). A variety of inference tech-
 9696 niques have been employed for this probabilistic model of relation extraction: Surdeanu
 9697 et al. (2012) use expectation maximization, Riedel et al. (2010) use sampling, and Hoff-
 9698 mann et al. (2011) use a custom graph-based algorithm. Expectation maximization and
 9699 sampling are surveyed in chapter 5, and are covered in more detail by Murphy (2012);
 9700 graph-based methods are surveyed by Mihalcea and Radev (2011).

9701 17.2.4 Open information extraction

9702 In classical relation extraction, the set of relations is defined in advance, using a **schema**.
 9703 The relation for any pair of entities can then be predicted using multi-class classification.
 9704 In **open information extraction** (OpenIE), a relation can be any triple of text. The example

Task	Relation ontology	Supervision
PropBank semantic role labeling	VerbNet	sentence
FrameNet semantic role labeling	FrameNet	sentence
Relation extraction	ACE, TAC, SemEval, etc	sentence
Slot filling	ACE, TAC, SemEval, etc	relation
Open Information Extraction	open	seed relations or patterns

Table 17.3: Various relation extraction tasks and their properties. VerbNet and FrameNet are described in chapter 13. ACE (Linguistic Data Consortium, 2005), TAC (McNamee and Dang, 2009), and SemEval (Hendrickx et al., 2009) refer to shared tasks, each of which involves an ontology of relation types.

9705 sentence (17.8a) instantiates several “relations” of this sort, e.g.,

- 9706 • (*mayor of, Maynard Jackson, Atlanta*),
- 9707 • (*elected, Maynard Jackson, mayor of Atlanta*),
- 9708 • (*elected in, Maynard Jackson, 1973*).

9709 Extracting such tuples can be viewed as a lightweight version of **semantic role labeling**
 9710 (chapter 13), with only two argument types: first slot and second slot. The task is gen-
 9711 erally evaluated on the relation level, rather than on the level of sentences: precision is
 9712 measured by the number of extracted relations that are accurate, and recall is measured
 9713 by the number of true relations that were successfully extracted. OpenIE systems are
 9714 trained from distant supervision or bootstrapping, rather than from labeled sentences.

9715 An early example is the TEXTRUNNER system (Banko et al., 2007), which identifies
 9716 relations with a set of handcrafted syntactic rules. The examples that are acquired from
 9717 the handcrafted rules are then used to train a classification model that uses part-of-speech
 9718 patterns as features. Finally, the relations that are extracted by the classifier are aggre-
 9719 gated, removing redundant relations and computing the number of times that each rela-
 9720 tion is mentioned in the corpus. TEXTRUNNER was the first in a series of systems that
 9721 performed increasingly accurate open relation extraction by incorporating more precise
 9722 linguistic features (Etzioni et al., 2011), distant supervision from Wikipedia infoboxes (Wu
 9723 and Weld, 2010), and better learning algorithms (Zhu et al., 2009).

9724 17.3 Events

9725 Relations link pairs of entities, but many real-world situations involve more than two enti-
 9726 ties. Consider again the example sentence (17.8a), which describes the event of an election,

9727 with four properties: the office (MAYOR), the district (ATLANTA), the date (1973), and the
9728 person elected (MAYNARD JACKSON, JR.). In **event detection**, a schema is provided for
9729 each event type (e.g., an election, a terrorist attack, or a chemical reaction), indicating all
9730 the possible properties of the event. The system is then required to fill in as many of these
9731 properties as possible (Doddington et al., 2004).

9732 Event detection systems generally involve a retrieval component (finding relevant
9733 documents and passages of text) and an extraction component (determining the proper-
9734 ties of the event based on the retrieved texts). Early approaches focused on finite-state pat-
9735 terns for identify event properties (Hobbs et al., 1997); such patterns can be automatically
9736 induced by searching for patterns that are especially likely to appear in documents that
9737 match the event query (Riloff, 1996). Contemporary approaches employ techniques that
9738 are similar to FrameNet semantic role labeling (§ 13.2), such as structured prediction over
9739 local and global features (Li et al., 2013) and bidirectional recurrent neural networks (Feng
9740 et al., 2016). These methods detect whether an event is described in a sentence, and if so,
9741 what are its properties.

9742 **Event coreference** Because multiple sentences may describe unique properties of a sin-
9743 gle event, **event coreference** is required to link event mentions across a single passage
9744 of text, or between passages (Humphreys et al., 1997). Bejan and Harabagiu (2014) de-
9745 fine event coreference as the task of identifying event mentions that share the same event
9746 participants (i.e., the slot-filling entities) and the same event properties (e.g., the time and
9747 location), within or across documents. Event coreference resolution can be performed us-
9748 ing supervised learning techniques in a similar way to entity coreference, as described
9749 in chapter 15: move left-to-right through the document, and use a classifier to decide
9750 whether to link each event reference to an existing cluster of coreferent events, or to cre-
9751 ate a new cluster (Ahn, 2006). Each clustering decision is based on the compatibility of
9752 features describing the participants and properties of the event. Due to the difficulty of
9753 annotating large amounts of data for entity coreference, unsupervised approaches are es-
9754 pecially desirable (Chen and Ji, 2009; Bejan and Harabagiu, 2014).

9755 **Relations between events** Just as entities are related to other entities, events may be
9756 related to other events: for example, the event of winning an election both *precedes* and
9757 *causes* the event of serving as mayor; moving to Atlanta *precedes* and *enables* the event of
9758 becoming mayor of Atlanta; moving from Dallas to Atlanta *prevents* the event of later be-
9759 coming mayor of Dallas. As these examples show, events may be related both temporally
9760 and causally. The **TimeML** annotation scheme specifies a set of six temporal relations
9761 between events (Pustejovsky et al., 2005), derived in part from **interval algebra** (Allen,
9762 1984). The TimeBank corpus provides TimeML annotations for 186 documents (Puste-
9763 jovsky et al., 2003). Methods for detecting these temporal relations combine supervised

	Positive (+)	Negative (-)	Underspecified (u)
Certain (CT)	Fact: CT+	Counterfact: CT-	Certain, but unknown: CTU
Probable (PR)	Probable: PR+	Not probable: PR-	(NA)
Possible (PS)	Possible: PS+	Not possible: PS-	(NA)
Underspecified (U)	(NA)	(NA)	Unknown or uncommitted: UU

Table 17.4: Table of factuality values from the FactBank corpus (Saurí and Pustejovsky, 2009). The entry (NA) indicates that this combination is not annotated.

9764 machine learning with temporal constraints, such as transitivity (e.g. Mani et al., 2006;
 9765 Chambers and Jurafsky, 2008).

9766 More recent annotation schemes and datasets combine temporal and causal relations (Mirza
 9767 et al., 2014; Dunietz et al., 2017): for example, the CaTeRS dataset includes annotations of
 9768 320 five-sentence short stories (Mostafazadeh et al., 2016). Abstracting still further, **pro-**
 9769 **cesses** are networks of causal relations between multiple events. A small dataset of bi-
 9770 ological processes is annotated in the ProcessBank dataset (Berant et al., 2014), with the
 9771 goal of supporting automatic question answering on scientific textbooks.

9772 17.4 Hedges, denials, and hypotheticals

9773 The methods described thus far apply to **propositions** about the way things are in the
 9774 real world. But natural language can also describe events and relations that are likely or
 9775 unlikely, possible or impossible, desired or feared. The following examples hint at the
 9776 scope of the problem (Prabhakaran et al., 2010):

- 9777 (17.9) a. GM will lay off workers.
 9778 b. A spokesman for GM said GM will lay off workers.
 9779 c. GM may lay off workers.
 9780 d. The politician claimed that GM will lay off workers.
 9781 e. Some wish GM would lay off workers.
 9782 f. Will GM lay off workers?
 9783 g. Many wonder whether GM will lay off workers.

9784 Accurate information extraction requires handling these **extra-propositional** aspects
 9785 of meaning, which are sometimes summarized under the terms **modality** and **negation**.⁷

⁷The classification of negation as extra-propositional is controversial: Packard et al. (2014) argue that negation is a “core part of compositionally constructed logical-form representations.” Negation is an element of the semantic parsing tasks discussed in chapter 12 and chapter 13 — for example, negation markers are

9786 Modality refers to expressions of the speaker's attitude towards her own statements, in-
9787 cluding "degree of certainty, reliability, subjectivity, sources of information, and perspec-
9788 tive" (Morante and Sporleder, 2012). Various systematizations of modality have been
9789 proposed (e.g., Palmer, 2001), including categories such as future, interrogative, imper-
9790 ative, conditional, and subjective. Information extraction is particularly concerned with
9791 negation and certainty. For example, Saurí and Pustejovsky (2009) link negation with
9792 a modal calculus of certainty, likelihood, and possibility, creating the two-dimensional
9793 schema shown in Table 17.4. This is the basis for the FactBank corpus, with annotations
9794 of the **factuality** of all sentences in 208 documents of news text.

9795 A related concept is **hedging**, in which speakers limit their commitment to a proposi-
9796 tion (Lakoff, 1973):

- 9797 (17.10) a. These results **suggest** that expression of c-jun, jun B and jun D genes **might** be
9798 involved in terminal granulocyte differentiation... (Morante and Daelemans,
9799 2009)
- 9800 b. A whale is **technically** a mammal (Lakoff, 1973)

9801 In the first example, the hedges *suggest* and *might* communicate uncertainty; in the second
9802 example, there is no uncertainty, but the hedge *technically* indicates that the evidence for
9803 the proposition will not fully meet the reader's expectations. Hedging has been studied
9804 extensively in scientific texts (Medlock and Briscoe, 2007; Morante and Daelemans, 2009),
9805 where the goal of large-scale extraction of scientific facts is obstructed by hedges and spec-
9806 ulation. Still another related aspect of modality is **evidentiality**, in which speakers mark
9807 the source of their information. In many languages, it is obligatory to mark evidentiality
9808 through affixes or particles (Aikhenvald, 2004); while evidentiality is not grammaticalized
9809 in English, authors are expected to express this information in contexts such as journal-
9810 ism (Kovach and Rosenstiel, 2014) and Wikipedia.⁸

9811 Methods for handling negation and modality generally include two phases:

- 9812 1. detecting negated or uncertain events;
9813 2. identifying **scope** of the negation or modal operator.

9814 A considerable body of work on negation has employed rule-based techniques such
9815 as regular expressions (Chapman et al., 2001) to detect negated events. Such techniques

treated as adjuncts in PropBank semantic role labeling. However, many of the relation extraction methods mentioned in this chapter do not handle negation directly. A further consideration is that negation interacts closely with aspects of modality that are generally not considered in propositional semantics, such as certainty and subjectivity.

⁸<https://en.wikipedia.org/wiki/Wikipedia:Verifiability>

match lexical cues (e.g., *Norwood was **not** elected Mayor*), while avoiding “double negatives” (e.g., *surely all this is **not without** meaning*). Supervised techniques involve classifiers over lexical and syntactic features (Uzuner et al., 2009) and sequence labeling (Prabhakaran et al., 2010).

The scope refers to the elements of the text whose propositional meaning is negated or modulated (Huddleston and Pullum, 2005), as elucidated in the following example from Morante and Sporleder (2012):

- (17.11) [After his habit he said] **nothing**, and after mine I asked no questions.
 After his habit he said nothing, and [after mine I asked] **no** [questions].

In this sentence, there are two negation cues (*nothing* and *no*). Each negates an event, indicated by the underlined verbs *said* and *asked*, and each occurs within a scope: *after his habit he said* and *after mine I asked* questions. Scope identification is typically formalized as sequence labeling problems, with each word token labeled as beginning, inside, or outside of a cue, focus, or scope span (see § 8.3). Conventional sequence labeling approaches can then be applied, using surface features as well as syntax (Veldal et al., 2012) and semantic analysis (Packard et al., 2014). Labeled datasets include the BioScope corpus of biomedical texts (Vincze et al., 2008) and a shared task dataset of detective stories by Arthur Conan Doyle (Morante and Blanco, 2012).

17.5 Question answering and machine reading

The victory of the Watson question-answering system against three top human players on the game show *Jeopardy!* was a landmark moment for natural language processing (Ferrucci et al., 2010). Game show questions are usually answered by **factoids**: entity names and short phrases.⁹ The task of factoid question answering is therefore closely related to information extraction, with the additional problem of accurately parsing the question.

17.5.1 Formal semantics

Semantic parsing is an effective method for question-answering in restricted domains such as questions about geography and airline reservations (Zettlemoyer and Collins, 2005), and has also been applied in “open-domain” settings such as question answering on Freebase (Berant et al., 2013) and biomedical research abstracts (Poon and Domingos, 2009). One approach is to convert the question into a lambda calculus expression that returns a boolean value: for example, the question *who is the mayor of the capital of Georgia?*

⁹The broader landscape of question answering includes “why” questions (*Why did Ahab continue to pursue the white whale?*), “how questions” (*How did Queequeg die?*), and requests for summaries (*What was Ishmael’s attitude towards organized religion?*). For more, see Hirschman and Gaizauskas (2001).

9847 would be converted to,

$$\lambda x. \exists y \text{ CAPITAL(GEORGIA, } y) \wedge \text{MAYOR}(y, x). \quad [17.22]$$

9848 This lambda expression can then be used to query an existing knowledge base, returning
9849 “true” for all entities that satisfy it.

9850 17.5.2 Machine reading

9851 Recent work has focused on answering questions about specific textual passages, similar
9852 to the reading comprehension examinations for young students (Hirschman et al., 1999).
9853 This task has come to be known as **machine reading**.

9854 Datasets

9855 The machine reading problem can be formulated in a number of different ways. The most
9856 important distinction is what form the answer should take.

9857 • **Multiple-choice question answering**, as in the MCTest dataset of stories (Richardson et al., 2013) and the New York Regents Science Exams (Clark, 2015). In MCTest,
9858 the answer is deducible from the text alone, while in the science exams, the system
9859 must make inferences using an existing model of the underlying scientific phenom-
9860 ena. Here is an example from MCTest:
9861

9862 (17.12) James the turtle was always getting into trouble. Sometimes he'd reach into
9863 the freezer and empty out all the food ...

9864 Q: What is the name of the trouble making turtle?
9865 (a) Fries
9866 (b) Pudding
9867 (c) James
9868 (d) Jane

9869 • **Cloze-style “fill in the blank” questions**, as in the CNN/Daily Mail comprehension
9870 task (Hermann et al., 2015), the Children’s Book Test (Hill et al., 2016), and the Who-
9871 did-What dataset (Onishi et al., 2016). In these tasks, the system must guess which
9872 word or entity completes a sentence, based on reading a passage of text. Here is an
9873 example from Who-did-What:

9874 (17.13) Q: Tottenham manager Juande Ramos has hinted he will allow ____ to leave
9875 if the Bulgaria striker makes it clear he is unhappy. (Onishi et al., 2016)

9876 The query sentence may be selected either from the story itself, or from an external
9877 summary. In either case, datasets can be created automatically by processing large

9878 quantities existing documents. An additional constraint is that that missing element
 9879 from the cloze must appear in the main passage of text: for example, in Who-did-
 9880 What, the candidates include all entities mentioned in the main passage. In the
 9881 CNN/Daily Mail dataset, each entity name is replaced by a unique identifier, e.g.,
 9882 ENTITY37. This ensures that correct answers can only be obtained by accurately
 9883 reading the text, and not from external knowledge about the entities.

9884 • **Extractive** question answering, in which the answer is drawn from the original text.
 9885 In WikiQA, answers are sentences (Yang et al., 2015). In the Stanford Question An-
 9886 swering Dataset (SQuAD), answers are words or short phrases (Rajpurkar et al.,
 9887 2016):

9888 (17.14) In metereology, precipitation is any product of the condensation of atmo-
 9889 spheric water vapor that falls under gravity.
 9890 Q: What causes precipitation to fall? A: gravity

9891 In both WikiQA and SQuAD, the original texts are Wikipedia articles, and the ques-
 9892 tions are generated by crowdworkers.

9893 Methods

9894 A baseline method is to search the text for sentences or short passages that overlap with
 9895 both the query and the candidate answer (Richardson et al., 2013). In example (17.12), this
 9896 baseline would select the correct answer, since *James* appears in a sentence that includes
 9897 the query terms *trouble* and *turtle*.

This baseline can be implemented as a neural architecture, using an **attention mechanism** (see § 18.3.1), which scores the similarity of the query to each part of the source text (Chen et al., 2016). The first step is to encode the passage $w^{(p)}$ and the query $w^{(q)}$, using two bidirectional LSTMs (§ 7.6).

$$\mathbf{h}^{(q)} = \text{BiLSTM}(\mathbf{w}^{(q)}; \Theta^{(q)}) \quad [17.23]$$

$$\mathbf{h}^{(p)} = \text{BiLSTM}(\mathbf{w}^{(p)}; \Theta^{(p)}). \quad [17.24]$$

The query is represented by vertically concatenating the final states of the left-to-right and right-to-left passes:

$$\mathbf{u} = [\overrightarrow{\mathbf{h}}^{(q)}_{M_q}; \overleftarrow{\mathbf{h}}^{(q)}_0]. \quad [17.25]$$

The attention vector is computed as a softmax over a vector of bilinear products, and the expected representation is computed by summing over attention values,

$$\tilde{\alpha}_m = (\mathbf{u}^{(q)})^\top \mathbf{W}_a \mathbf{h}_m^{(p)} \quad [17.26]$$

$$\boldsymbol{\alpha} = \text{SoftMax}(\tilde{\boldsymbol{\alpha}}) \quad [17.27]$$

$$\mathbf{o} = \sum_{m=1}^M \alpha_m \mathbf{h}_m^{(p)}. \quad [17.28]$$

Each candidate answer c is represented by a vector \mathbf{x}_c . Assuming the candidate answers are spans from the original text, these vectors can be set equal to the corresponding element in $\mathbf{h}^{(p)}$. The score for each candidate answer a is computed by the inner product,

$$\hat{c} = \underset{c}{\operatorname{argmax}} \mathbf{o} \cdot \mathbf{x}_c. \quad [17.29]$$

9898 This architecture can be trained end-to-end from a loss based on the log-likelihood of the
 9899 correct answer. A number of related architectures have been proposed (e.g., Hermann
 9900 et al., 2015; Kadlec et al., 2016; Dhingra et al., 2017; Cui et al., 2017), and these methods are
 9901 surveyed by Wang et al. (2017).

9902 Additional resources

9903 The field of information extraction is surveyed in course notes by Grishman (2012), and
 9904 more recently in a short survey paper (Grishman, 2015). Shen et al. (2015) survey the task
 9905 of entity linking, and Ji and Grishman (2011) survey work on knowledge base popula-
 9906 tion. This chapter’s discussion of non-propositional meaning was strongly influenced by
 9907 Morante and Sporleder (2012), who introduced a special issue of the journal *Computational
 9908 Linguistics* dedicated to recent work on modality and negation.

9909 Exercises

- 9910 1. Go to the Wikipedia page for your favorite movie. For each record in the info box
 9911 (e.g., *Screenplay by: Stanley Kubrick*), report whether there is a sentence in the ar-
 9912 ticle containing both the field and value (e.g., *The screenplay was written by Stanley
 9913 Kubrick*). If not, is there a sentence in the article containing just the value? (For
 9914 records with more than one value, just use the first value.)
- 9915 2. Building on your answer in the previous question, report the dependency path be-
 9916 tween the head words of the field and value for at least three records.
- 9917 3. Consider the following heuristic for entity linking:

- 9918 • Among all entities that have the same type as the mention (e.g., LOC, PER),
 9919 choose the one whose name has the lowest edit distance from the mention.
 9920 • If more than one entity has the right type and the lowest edit distance from the
 9921 mention, choose the most popular one.
 9922 • If no candidate entity has the right type, choose NIL.

Now suppose you have the following feature function:

$$\mathbf{f}(y, \mathbf{x}) = [\text{edit-dist}(\text{name}(y), \mathbf{x}), \text{same-type}(y, \mathbf{x}), \text{popularity}(y), \delta(y = \text{NIL})]$$

9923 Design a set of ranking weights θ that match the heuristic. You may assume that
 9924 edit distance and popularity are always in the range [0, 100], and that the NIL entity
 9925 has values of zero for all features except $\delta(y = \text{NIL})$.

9926 4. Now consider another heuristic:

- 9927 • Among all candidate entities that have edit distance zero from the mention,
 9928 and are the right type, choose the most popular one.
 9929 • If no entity has edit distance zero from the mention, choose the one with the
 9930 right type that is most popular, regardless of edit distance.
 9931 • If no entity has the right type, choose NIL.

9932 Using the same features and assumptions from the previous problem, prove that
 9933 there is no set of weights that could implement this heuristic. Then show that the
 9934 heuristic can be implemented by adding a single feature. Your new feature should
 9935 consider only the edit distance.

9936 5. Download the Reuters corpus in NLTK, and iterate over the tokens in the corpus:

```
9937   import nltk
9938   nltk.corpus.download('reuters')
9939   from nltk.corpus import reuters
9940   for word in reuters.words():
9941     #your code here
```

- 9942 a) Apply the pattern , such as to obtain candidates for the IS-A relation,
 9943 e.g. IS-A(ROMANIA, COUNTRY). What are three pairs that this method identi-
 9944 fies correctly? What are three different pairs that it gets wrong?
 9945 b) Design a pattern for the PRESIDENT relation, e.g. PRESIDENT(PHILIPPINES, CORAZON AQUINO)
 9946 In this case, you may want to augment your pattern matcher with the ability
 9947 to match multiple token wildcards, perhaps using case information to detect
 9948 proper names. Again, list three correct

- 9949 c) Preprocess the Reuters data by running a named entity recognizer, replacing
 9950 tokens with named entity spans when applicable — e.g., your pattern can now
 9951 match on *the United States* if the NER system tags it. Apply your PRESIDENT
 9952 matcher to this preprocessed data. Does the accuracy improve? Compare 20
 9953 randomly-selected pairs from this pattern and the one you designed in the pre-
 9954 vious part.
- 9955 6. Using the same NLTK Reuters corpus, apply distant supervision to build a training
 9956 set for detecting the relation between nations and their capitals. Start with the fol-
 9957 lowing known relations: (JAPAN, TOKYO), (FRANCE, PARIS), (ITALY, ROME). How
 9958 many positive and negative examples are you able to extract?
- 9959 7. Represent the dependency path $\mathbf{x}^{(i)}$ as a sequence of words and dependency arcs
 9960 of length M_i , ignoring the endpoints of the path. In example 1 of Table 17.2, the
 9961 dependency path is,

$$\mathbf{x}^{(1)} = (\xleftarrow[\text{NSUBJ}]{} \text{traveled}, \xrightarrow[\text{OBL}]{}) \quad [17.30]$$

9962 If $x_m^{(i)}$ is a word, then let $\text{pos}(x_m^{(i)})$ be its part-of-speech, using the tagset defined in
 9963 chapter 8.

We can define the following kernel function over pairs of dependency paths (Bunescu and Mooney, 2005):

$$\kappa(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \begin{cases} 0, & M_i \neq M_j \\ \prod_{m=1}^{M_i} c(x_m^{(i)}, x_m^{(j)}), & M_i = M_j \end{cases}$$

$$c(x_m^{(i)}, x_m^{(j)}) = \begin{cases} 2, & x_m^{(i)} = x_m^{(j)} \\ 1, & x_m^{(i)} \neq x_m^{(j)} \text{ and } \text{pos}(x_m^{(i)}) = \text{pos}(x_m^{(j)}) \\ 0, & \text{otherwise.} \end{cases}$$

9964 Using this kernel function, compute the kernel similarities of example 1 from Ta-
 9965 ble 17.2 with the other five examples.

8. Continuing from the previous problem, suppose that the instances have the following labels:

$$y_2 = 1, y_3 = -1, y_4 = -1, y_5 = 1, y_6 = 1 \quad [17.31]$$

9966 Equation 17.13 defines a kernel-based classification in terms of parameters α and
 9967 b . Using the above labels for y_2, \dots, y_6 , identify the values of α and b under which
 9968 $\hat{y}_1 = 1$. Remember the constraint that $\alpha_i \geq 0$ for all i .

9969 9. Consider the neural QA system described in § 17.5.2, but restrict the set of candidate
 9970 answers to words in the passage, and set each candidate answer embedding x equal
 9971 to the vector $\mathbf{h}_m^{(p)}$, representing token m in the passage, so that $\hat{m} = \operatorname{argmax}_m \mathbf{o} \cdot \mathbf{h}_m^{(p)}$.
 9972 Suppose the system selects answer \hat{m} , but the correct answer is m^* . Consider the
 9973 gradient of the margin loss with respect to the attention:

- 9974 a) Prove that $\frac{\partial \ell}{\partial \alpha_{\hat{m}}} \geq \frac{\partial \ell}{\partial \alpha_{m^*}}$.
- 9975 b) Assuming that $\|\mathbf{h}_{\hat{m}}\| = \|\mathbf{h}_{m^*}\|$, prove that $\frac{\partial \ell}{\partial \alpha_{\hat{m}}} \geq 0$ and $\frac{\partial \ell}{\partial \alpha_{m^*}} \leq 0$. Explain in
 9976 words what this means about how the attention is expected to change after a
 9977 gradient-based update.

9978 Chapter 18

9979 Machine translation

9980 Machine translation (MT) is one of the “holy grail” problems in artificial intelligence,
9981 with the potential to transform society by facilitating communication between people
9982 anywhere in the world. As a result, MT has received significant attention and funding
9983 since the early 1950s. However, it has proved remarkably challenging, and while there
9984 has been substantial progress towards usable MT systems — especially for high-resource
9985 language pairs like English-French — we are still far from translation systems that match
9986 the nuance and depth of human translations.

9987 18.1 Machine translation as a task

9988 Machine translation can be formulated as an optimization problem:

$$\hat{\mathbf{w}}^{(t)} = \underset{\mathbf{w}^{(t)}}{\operatorname{argmax}} \Psi(\mathbf{w}^{(s)}, \mathbf{w}^{(t)}), \quad [18.1]$$

9989 where $\mathbf{w}^{(s)}$ is a sentence in a **source** language, $\mathbf{w}^{(t)}$ is a sentence in the **target language**,
9990 and Ψ is a scoring function. As usual, this formalism requires two components: a decod-
9991 ing algorithm for computing $\hat{\mathbf{w}}^{(t)}$, and a learning algorithm for estimating the parameters
9992 of the scoring function Ψ .

9993 Decoding is difficult for machine translation because of the huge space of possible
9994 translations. We have faced large label spaces before: for example, in sequence labeling,
9995 the set of possible label sequences is exponential in the length of the input. In these cases,
9996 it was possible to search the space quickly by introducing locality assumptions: for ex-
9997 ample, that each tag depends only on its predecessor, or that each production depends
9998 only on its parent. In machine translation, no such locality assumptions seem possible:
9999 human translators reword, reorder, and rearrange words; they replace single words with
10000 multi-word phrases, and vice versa. This flexibility means that in even relatively simple

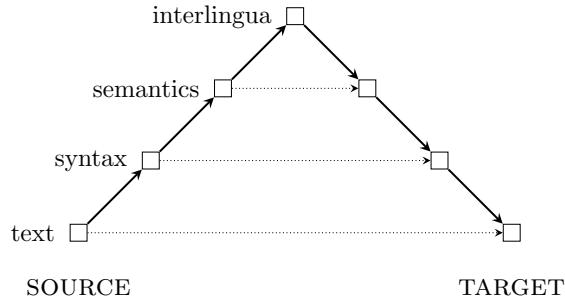


Figure 18.1: The Vauquois Pyramid

10001 translation models, decoding is NP-hard (Knight, 1999). Approaches for dealing with this
 10002 complexity are described in § 18.4.

Estimating translation models is difficult as well. Labeled translation data usually comes in the form parallel sentences, e.g.,

$$\begin{aligned} \mathbf{w}^{(s)} &= A \text{ Vinay le gusta las manzanas.} \\ \mathbf{w}^{(t)} &= \text{Vinay likes apples.} \end{aligned}$$

10003 A useful feature function would note the translation pairs (*gusta, likes*), (*manzanas, apples*),
 10004 and even (*Vinay, Vinay*). But this word-to-word **alignment** is not given in the data. One
 10005 solution is to treat this alignment as a **latent variable**; this is the approach taken by clas-
 10006 sical **statistical machine translation** (SMT) systems, described in § 18.2. Another solution
 10007 is to model the relationship between $\mathbf{w}^{(t)}$ and $\mathbf{w}^{(s)}$ through a more complex and expres-
 10008 sive function; this is the approach taken by **neural machine translation** (NMT) systems,
 10009 described in § 18.3.

10010 The **Vauquois Pyramid** is a theory of how translation should be done. At the lowest
 10011 level, the translation system operates on individual words, but the horizontal distance
 10012 at this level is large, because languages express ideas differently. If we can move up the
 10013 triangle to syntactic structure, the distance for translation is reduced; we then need only
 10014 produce target-language text from the syntactic representation, which can be as simple
 10015 as reading off a tree. Further up the triangle lies semantics; translating between semantic
 10016 representations should be easier still, but mapping between semantics and surface text is a
 10017 difficult, unsolved problem. At the top of the triangle is **interlingua**, a semantic represen-
 10018 tation that is so generic that it is identical across all human languages. Philosophers de-
 10019 bate whether such a thing as interlingua is really possible (e.g., Derrida, 1985). While the
 10020 first-order logic representations discussed in chapter 12 might be thought to be language
 10021 independent, they are built on an inventory of predicates that are suspiciously similar to
 10022 English words (Nirenburg and Wilks, 2001). Nonetheless, the idea of linking translation

	Adequate?	Fluent?
<i>To Vinay it like Python</i>	yes	no
<i>Vinay debugs memory leaks</i>	no	yes
<i>Vinay likes Python</i>	yes	yes

Table 18.1: Adequacy and fluency for translations of the Spanish sentence *A Vinay le gusta Python*.

and semantic understanding may still be a promising path, if the resulting translations better preserve the meaning of the original text.

18.1.1 Evaluating translations

There are two main criteria for a translation, summarized in Table 18.1.

- **Adequacy:** The translation $w^{(t)}$ should adequately reflect the linguistic content of $w^{(s)}$. For example, if $w^{(s)} = A Vinay le gusta Python$, the reference translation is $w^{(t)} = Vinay likes Python$. However, the **gloss**, or word-for-word translation $w^{(t)} = To Vinay it like Python$ is also considered adequate because it contains all the relevant content. The output $w^{(t)} = Vinay debugs memory leaks$ is not adequate.
- **Fluency:** The translation $w^{(t)}$ should read like fluent text in the target language. By this criterion, the gloss $w^{(t)} = To Vinay it like Python$ will score poorly, and $w^{(t)} = Vinay debugs memory leaks$ will be preferred.

Automated evaluations of machine translations typically merge both of these criteria, by comparing the system translation with one or more **reference translations**, produced by professional human translators. The most popular quantitative metric is **BLEU** (bilingual evaluation understudy; Papineni et al., 2002), which is based on n -gram precision: what fraction of n -grams in the system translation appear in the reference? Specifically, for each n -gram length, the precision is defined as,

$$p_n = \frac{\text{number of } n\text{-grams appearing in both reference and hypothesis translations}}{\text{number of } n\text{-grams appearing in the hypothesis translation}}. \quad [18.2]$$

The n -gram precisions for three hypothesis translations are shown in Figure 18.2.

The BLEU score is then based on the average, $\exp \frac{1}{N} \sum_{n=1}^N \log p_n$. Two modifications of Equation 18.2 are necessary: (1) to avoid computing $\log 0$, all precisions are smoothed to ensure that they are positive; (2) each n -gram in the reference can be used at most once, so that *to to to to* does not achieve $p_1 = 1$ against the reference *to be or not to be*. Furthermore, precision-based metrics are biased in favor of short translations, which

	Translation	p_1	p_2	p_3	p_4	BP	BLEU
<i>Reference</i>	<i>Vinay likes programming in Python</i>						
<i>Sys1</i>	<i>To Vinay it like to program Python</i>	$\frac{2}{7}$	0	0	0	1	.21
<i>Sys2</i>	<i>Vinay likes Python</i>	$\frac{3}{3}$	$\frac{1}{2}$	0	0	.51	.33
<i>Sys3</i>	<i>Vinay likes programming in his pajamas</i>	$\frac{4}{6}$	$\frac{3}{5}$	$\frac{2}{4}$	$\frac{1}{3}$	1	.76

Figure 18.2: A reference translation and three system outputs. For each output, p_n indicates the precision at each n -gram, and BP indicates the brevity penalty.

10047 can achieve high scores by minimizing the denominator in [18.2]. To avoid this issue, a
 10048 **brevity penalty** is applied to translations that are shorter than the reference. This penalty
 10049 is indicated as “BP” in Figure 18.2.

10050 Automated metrics like BLEU have been validated by correlation with human judg-
 10051 ments of translation quality. Nonetheless, it is not difficult to construct examples in which
 10052 the BLEU score is high, yet the translation is disfluent or carries a completely different
 10053 meaning from the original. To give just one example, consider the problem of translating
 10054 pronouns. Because pronouns refer to specific entities, a single incorrect pronoun can obl-
 10055 erate the semantics of the original sentence. Existing state-of-the-art systems generally
 10056 do not attempt the reasoning necessary to correctly resolve pronominal anaphora (Hard-
 10057 meier, 2012). Despite the importance of pronouns for semantics, they have a marginal
 10058 impact on BLEU, which may help to explain why existing systems do not make a greater
 10059 effort to translate them correctly.

10060 **Fairness and bias** The problem of pronoun translation intersects with issues of fairness
 10061 and bias. In many languages, such as Turkish, the third person singular pronoun is gender
 10062 neutral. Today’s state-of-the-art systems produce the following Turkish-English transla-
 10063 tions (Caliskan et al., 2017):

- 10064 (18.1) *O bir doktor.*
 He is a doctor.
 10065 (18.2) *O bir hemşire.*
 She is a nurse.

10066 The same problem arises for other professions that have stereotypical genders, such as
 10067 engineers, soldiers, and teachers, and for other languages that have gender-neutral pro-
 10068 nouns. This bias was not directly programmed into the translation model; it arises from
 10069 statistical tendencies in existing datasets. This highlights a general problem with data-
 10070 driven approaches, which can perpetuate biases that negatively impact disadvantaged

10071 groups. Worse, machine learning can *amplify* biases in data (Bolukbasi et al., 2016): if a
10072 dataset has even a slight tendency towards men as doctors, the resulting translation model
10073 may produce translations in which doctors are always *he*, and nurses are always *she*.

10074 **Other metrics** A range of other automated metrics have been proposed for machine
10075 translation. One potential weakness of BLEU is that it only measures precision; METEOR
10076 is a weighted *F*-MEASURE, which is a combination of recall and precision (see § 4.4.1).
10077 **Translation Error Rate (TER)** computes the string **edit distance** (see § 9.1.4) between the
10078 reference and the hypothesis (Snover et al., 2006). For language pairs like English and
10079 Japanese, there are substantial differences in word order, and word order errors are not
10080 sufficiently captured by *n*-gram based metrics. The **RIBES** metric applies rank correla-
10081 tion to measure the similarity in word order between the system and reference transla-
10082 tions (Isozaki et al., 2010).

10083 18.1.2 Data

10084 Data-driven approaches to machine translation rely primarily on **parallel corpora**, which
10085 are translations at the sentence level. Early work focused on government records, in which
10086 fine-grained official translations are often required. For example, the IBM translation sys-
10087 tems were based on the proceedings of the Canadian Parliament, called **Hansards**, which
10088 are recorded in English and French (Brown et al., 1990). The growth of the European
10089 Union led to the development of the **EuroParl corpus**, which spans 21 European lan-
10090 guages (Koehn, 2005). While these datasets helped to launch the field of machine transla-
10091 tion, they are restricted to narrow domains and a formal speaking style, limiting their ap-
10092 plicability to other types of text. As more resources are committed to machine translation,
10093 new translation datasets have been commissioned. This has broadened the scope of avail-
10094 able data to news,¹ movie subtitles,² social media (Ling et al., 2013), dialogues (Fordyce,
10095 2007), TED talks (Paul et al., 2010), and scientific research articles (Nakazawa et al., 2016).

10096 Despite this growing set of resources, the main bottleneck in machine translation data
10097 is the need for parallel corpora that are aligned at the sentence level. Many languages have
10098 sizable parallel corpora with some high-resource language, but not with each other. The
10099 high-resource language can then be used as a “pivot” or “bridge” (Boitet, 1988; Utiyama
10100 and Isahara, 2007): for example, De Gispert and Marino (2006) use Spanish as a bridge for
10101 translation between Catalan and English. For most of the 6000 languages spoken today,
10102 the only source of translation data remains the Judeo-Christian Bible (Resnik et al., 1999).
10103 While relatively small, at less than a million tokens, the Bible has been translated into
10104 more than 2000 languages, far outpacing any other corpus. Some research has explored

¹https://catalog.ldc.upenn.edu/LDC2010T10_translation-task.html <http://www.statmt.org/wmt15/>

²<http://opus.nlpl.eu/>

10105 the possibility of automatically identifying parallel sentence pairs from unaligned parallel
 10106 texts, such as web pages and Wikipedia articles (Kilgarriff and Grefenstette, 2003; Resnik
 10107 and Smith, 2003; Adafre and De Rijke, 2006). Another approach is to create large parallel
 10108 corpora through crowdsourcing (Zaidan and Callison-Burch, 2011).

10109 18.2 Statistical machine translation

10110 The previous section introduced adequacy and fluency as the two main criteria for ma-
 10111 chine translation. A natural modeling approach is to represent them with separate scores,

$$\Psi(\mathbf{w}^{(s)}, \mathbf{w}^{(t)}) = \Psi_A(\mathbf{w}^{(s)}, \mathbf{w}^{(t)}) + \Psi_F(\mathbf{w}^{(t)}). \quad [18.3]$$

10112 The fluency score Ψ_F need not even consider the source sentence; it only judges $\mathbf{w}^{(t)}$ on
 10113 whether it is fluent in the target language. This decomposition is advantageous because
 10114 it makes it possible to estimate the two scoring functions on separate data. While the
 10115 adequacy model must be estimated from aligned sentences — which are relatively expen-
 10116 sive and rare — the fluency model can be estimated from monolingual text in the target
 10117 language. Large monolingual corpora are now available in many languages, thanks to
 10118 resources such as Wikipedia.

An elegant justification of the decomposition in Equation 18.3 is provided by the **noisy channel model**, in which each scoring function is a log probability:

$$\Psi_A(\mathbf{w}^{(s)}, \mathbf{w}^{(t)}) \triangleq \log p_{S|T}(\mathbf{w}^{(s)} | \mathbf{w}^{(t)}) \quad [18.4]$$

$$\Psi_F(\mathbf{w}^{(t)}) \triangleq \log p_T(\mathbf{w}^{(t)}) \quad [18.5]$$

$$\Psi(\mathbf{w}^{(s)}, \mathbf{w}^{(t)}) = \log p_{S|T}(\mathbf{w}^{(s)} | \mathbf{w}^{(t)}) + \log p_T(\mathbf{w}^{(t)}) = \log p_{S,T}(\mathbf{w}^{(s)}, \mathbf{w}^{(t)}). \quad [18.6]$$

10119 By setting the scoring functions equal to the logarithms of the prior and likelihood, their
 10120 sum is equal to $\log p_{S,T}$, which is the logarithm of the joint probability of the source and
 10121 target. The sentence $\hat{\mathbf{w}}^{(t)}$ that maximizes this joint probability is also the maximizer of the
 10122 conditional probability $p_{T|S}$, making it the most likely target language sentence, condi-
 10123 tioned on the source.

10124 The noisy channel model can be justified by a generative story. The target text is orig-
 10125 inally generated from a probability model p_T . It is then encoded in a “noisy channel”
 10126 $p_{S|T}$, which converts it to a string in the source language. In decoding, we apply Bayes’
 10127 rule to recover the string $\mathbf{w}^{(t)}$ that is maximally likely under the conditional probability
 10128 $p_{T|S}$. Under this interpretation, the target probability p_T is just a language model, and
 10129 can be estimated using any of the techniques from chapter 6. The only remaining learning
 10130 problem is to estimate the translation model $p_{S|T}$.

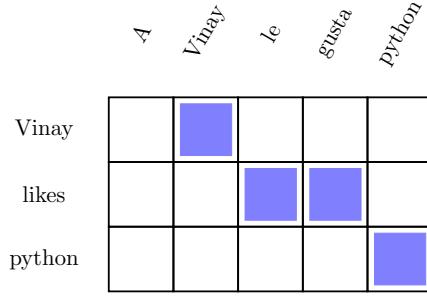


Figure 18.3: An example word-to-word alignment

18.2.1 Statistical translation modeling

The simplest decomposition of the translation model is word-to-word: each word in the source should be aligned to a word in the translation. This approach presupposes an **alignment** $\mathcal{A}(\mathbf{w}^{(s)}, \mathbf{w}^{(t)})$, which contains a list of pairs of source and target tokens. For example, given $\mathbf{w}^{(s)} = A\ Vinay\ le\ gusta\ Python$ and $\mathbf{w}^{(t)} = Vinay\ likes\ Python$, one possible word-to-word alignment is,

$$\mathcal{A}(\mathbf{w}^{(s)}, \mathbf{w}^{(t)}) = \{(A, \emptyset), (Vinay, Vinay), (le, likes), (gusta, likes), (Python, Python)\}. \quad [18.7]$$

This alignment is shown in Figure 18.3. Another, less promising, alignment is:

$$\mathcal{A}(\mathbf{w}^{(s)}, \mathbf{w}^{(t)}) = \{(A, Vinay), (Vinay, likes), (le, Python), (gusta, \emptyset), (Python, \emptyset)\}. \quad [18.8]$$

Each alignment contains exactly one tuple for each word in the *source*, which serves to explain how the source word could be translated from the target, as required by the translation probability $p_{S|T}$. If no appropriate word in the target can be identified for a source word, it is aligned to \emptyset — as is the case for the Spanish function word *a* in the example, which glosses to the English word *to*. Words in the target can align with multiple words in the source, so that the target word *likes* can align to both *le* and *gusta* in the source.

The joint probability of the alignment and the translation can be defined conveniently as,

$$p(\mathbf{w}^{(s)}, \mathcal{A} | \mathbf{w}^{(t)}) = \prod_{m=1}^{M^{(s)}} p(w_m^{(s)}, a_m | w_{a_m}^{(t)}, m, M^{(s)}, M^{(t)}) \quad [18.9]$$

$$= \prod_{m=1}^{M^{(s)}} p(a_m | m, M^{(s)}, M^{(t)}) \times p(w_m^{(s)} | w_{a_m}^{(t)}). \quad [18.10]$$

This probability model makes two key assumptions:

- 10145 • The alignment probability factors across tokens,

$$p(\mathcal{A} \mid \mathbf{w}^{(s)}, \mathbf{w}^{(t)}) = \prod_{m=1}^{M^{(s)}} p(a_m \mid m, M^{(s)}, M^{(t)}). \quad [18.11]$$

10146 This means that each alignment decision is independent of the others, and depends
 10147 only on the index m , and the sentence lengths $M^{(s)}$ and $M^{(t)}$.

- 10148 • The translation probability also factors across tokens,

$$p(\mathbf{w}^{(s)} \mid \mathbf{w}^{(t)}, \mathcal{A}) = \prod_{m=1}^{M^{(s)}} p(w_m^{(s)} \mid w_{a_m}^{(t)}), \quad [18.12]$$

10149 so that each word in $\mathbf{w}^{(s)}$ depends only on its aligned word in $\mathbf{w}^{(t)}$. This means that
 10150 translation is word-to-word, ignoring context. The hope is that the target language
 10151 model $p(\mathbf{w}^{(t)})$ will correct any disfluencies that arise from word-to-word translation.

To translate with such a model, we could sum or max over all possible alignments,

$$p(\mathbf{w}^{(s)}, \mathbf{w}^{(t)}) = \sum_{\mathcal{A}} p(\mathbf{w}^{(s)}, \mathbf{w}^{(t)}, \mathcal{A}) \quad [18.13]$$

$$= p(\mathbf{w}^{(t)}) \sum_{\mathcal{A}} p(\mathcal{A}) \times p(\mathbf{w}^{(s)} \mid \mathbf{w}^{(t)}, \mathcal{A}) \quad [18.14]$$

$$\geq p(\mathbf{w}^{(t)}) \max_{\mathcal{A}} p(\mathcal{A}) \times p(\mathbf{w}^{(s)} \mid \mathbf{w}^{(t)}, \mathcal{A}). \quad [18.15]$$

The term $p(\mathcal{A})$ defines the prior probability over alignments. A series of alignment models with increasingly relaxed independence assumptions was developed by researchers at IBM in the 1980s and 1990s, known as IBM Models 1-6 (Och and Ney, 2003). IBM Model 1 makes the strongest independence assumption:

$$p(a_m \mid m, M^{(s)}, M^{(t)}) = \frac{1}{M^{(t)}}. \quad [18.16]$$

10152 In this model, every alignment is equally likely. This is almost surely wrong, but it re-
 10153 sults in a convex learning objective, yielding a good initialization for the more complex
 10154 alignment models (Brown et al., 1993; Koehn, 2009).

10155 18.2.2 Estimation

10156 Let us define the parameter $\theta_{u \rightarrow v}$ as the probability of translating target word u to source
 10157 word v . If word-to-word alignments were annotated, these probabilities could be com-
 10158 puted from relative frequencies,

$$\hat{\theta}_{u \rightarrow v} = \frac{\text{count}(u, v)}{\text{count}(u)}, \quad [18.17]$$

10159 where $\text{count}(u, v)$ is the count of instances in which word v was aligned to word u in
 10160 the training set, and $\text{count}(u)$ is the total count of the target word u . The smoothing
 10161 techniques mentioned in chapter 6 can help to reduce the variance of these probability
 10162 estimates.

10163 Conversely, if we had an accurate translation model, we could estimate the likelihood
 10164 of each alignment decision,

$$q_m(a_m \mid \mathbf{w}^{(s)}, \mathbf{w}^{(t)}) \propto p(a_m \mid m, M^{(s)}, M^{(t)}) \times p(w_m^{(s)} \mid w_{a_m}^{(t)}), \quad [18.18]$$

where $q_m(a_m \mid \mathbf{w}^{(s)}, \mathbf{w}^{(t)})$ is a measure of our confidence in aligning source word $w_m^{(s)}$
 to target word $w_{a_m}^{(t)}$. The relative frequencies could then be computed from the *expected
 counts*,

$$\hat{\theta}_{u \rightarrow v} = \frac{E_q [\text{count}(u, v)]}{\text{count}(u)} \quad [18.19]$$

$$E_q [\text{count}(u, v)] = \sum_m q_m(a_m \mid \mathbf{w}^{(s)}, \mathbf{w}^{(t)}) \times \delta(w_m^{(s)} = v) \times \delta(w_{a_m}^{(t)} = u). \quad [18.20]$$

10165 The **expectation-maximization** (EM) algorithm proceeds by iteratively updating q_m
 10166 and $\hat{\Theta}$. The algorithm is described in general form in chapter 5. For statistical machine
 10167 translation, the steps of the algorithm are:

- 10168 1. **E-step:** Update beliefs about word alignment using Equation 18.18.
 10169 2. **M-step:** Update the translation model using Equations 18.19 and 18.20.

10170 As discussed in chapter 5, the expectation maximization algorithm is guaranteed to con-
 10171 verge, but not to a global optimum. However, for IBM Model 1, it can be shown that EM
 10172 optimizes a convex objective, and global optimality is guaranteed. For this reason, IBM
 10173 Model 1 is often used as an initialization for more complex alignment models. For more
 10174 detail, see Koehn (2009).

10175 18.2.3 Phrase-based translation

10176 Real translations are not word-to-word substitutions. One reason is that many multiword
 10177 expressions are not translated literally, as shown in this example from French:

- 10178 (18.3) *Nous allons prendre un verre*
 We will take a glass
 10179 We'll have a drink

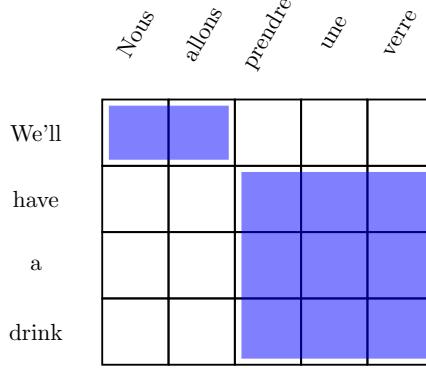


Figure 18.4: A phrase-based alignment between French and English, corresponding to example (18.3)

10180 The line *we will take a glass* is the word-for-word gloss of the French sentence; the translation
 10181 *we'll have a drink* is shown on the third line. Such examples are difficult for word-to-
 10182 word translation models, since they require translating *prendre* to *have* and *verre* to *drink*.
 10183 These translations are only correct in the context of these specific phrases.

Phrase-based translation generalizes on word-based models by building translation tables and alignments between multiword spans. (These “phrases” are not necessarily syntactic constituents like the noun phrases and verb phrases described in chapters 9 and 10.) The generalization from word-based translation is surprisingly straightforward: the translation tables can now condition on multi-word units, and can assign probabilities to multi-word units; alignments are mappings from spans to spans, $((i, j), (k, \ell))$, so that

$$p(\mathbf{w}^{(s)} | \mathbf{w}^{(t)}, \mathcal{A}) = \prod_{((i, j), (k, \ell)) \in \mathcal{A}} p_{w^{(s)}|w^{(t)}}(\{w_{i+1}^{(s)}, w_{i+2}^{(s)}, \dots, w_j^{(s)}\} | \{w_{k+1}^{(t)}, w_{k+2}^{(t)}, \dots, w_\ell^{(t)}\}). \quad [18.21]$$

10184 The phrase alignment $((i, j), (k, \ell))$ indicates that the span $\mathbf{w}_{i+1:j}^{(s)}$ is the translation of the
 10185 span $\mathbf{w}_{k+1:\ell}^{(t)}$. An example phrasal alignment is shown in Figure 18.4. Note that the align-
 10186 ment set \mathcal{A} is required to cover all of the tokens in the source, just as in word-based trans-
 10187 lation. The probability model $p_{w^{(s)}|w^{(t)}}$ must now include translations for all phrase pairs,
 10188 which can be learned from expectation-maximization just as in word-based statistical ma-
 10189 chine translation.

10190 **18.2.4 *Syntax-based translation**

10191 The Vauquois Pyramid (Figure 18.1) suggests that translation might be easier if we take a
 10192 higher-level view. One possibility is to incorporate the syntactic structure of the source,
 10193 the target, or both. This is particularly promising for language pairs that consistent syn-
 10194 tactic differences. For example, English adjectives almost always precede the nouns that
 10195 they modify, while in Romance languages such as French and Spanish, the adjective often
 10196 follows the noun: thus, *angry fish* would translate to *pez (fish) enojado (angry)* in Spanish.
 10197 In word-to-word translation, these reorderings cause the alignment model to be overly
 10198 permissive. It is not that the order of *any* pair of English words can be reversed when
 10199 translating into Spanish, but only adjectives and nouns within a noun phrase. Similar
 10200 issues arise when translating between verb-final languages such as Japanese (in which
 10201 verbs usually follow the subject and object), verb-initial languages like Tagalog and clas-
 10202 sical Arabic, and verb-medial languages such as English.

10203 An elegant solution is to link parsing and translation in a **synchronous context-free**
 10204 **grammar** (SCFG; Chiang, 2007).³ An SCFG is a set of productions of the form $X \rightarrow (\alpha, \beta, \sim)$,
 10205 where X is a non-terminal, α and β are sequences of terminals or non-terminals, and \sim
 10206 is a one-to-one alignment of items in α with items in β . English-Spanish adjective-noun
 10207 ordering can be handled by a set of synchronous productions, e.g.,

$$\text{NP} \rightarrow (\text{DET}_1 \text{NN}_2 \text{JJ}_3, \quad \text{DET}_1 \text{JJ}_3 \text{NN}_2), \quad [18.22]$$

10208 with subscripts indicating the alignment between the Spanish (left) and English (right)
 10209 parts of the right-hand side. Terminal productions yield translation pairs,

$$\text{JJ} \rightarrow (\text{enojado}_1, \quad \text{angry}_1). \quad [18.23]$$

10210 A synchronous derivation begins with the start symbol S , and derives a pair of sequences
 10211 of terminal symbols.

10212 Given an SCFG in which each production yields at most two symbols in each lan-
 10213 guage (Chomsky Normal Form; see § 9.2.1), a sentence can be parsed using only the CKY
 10214 algorithm (chapter 10). The resulting derivation also includes productions in the other
 10215 language, all the way down to the surface form. Therefore, SCFGs make translation very
 10216 similar to parsing. In a weighted SCFG, the log probability $\log p_{S|T}$ can be computed from
 10217 the sum of the log-probabilities of the productions. However, combining SCFGs with a
 10218 target language model is computationally expensive, necessitating approximate search
 10219 algorithms (Huang and Chiang, 2007).

10220 Synchronous context-free grammars are an example of **tree-to-tree translation**, be-
 10221 cause they model the syntactic structure of both the target and source language. In **string-**
 10222 **to-tree translation**, string elements are translated into constituent tree fragments, which

³Earlier approaches to syntactic machine translation includes syntax-driven transduction (Lewis II and Stearns, 1968) and stochastic inversion grammars (Wu, 1997).

10223 are then assembled into a translation (Yamada and Knight, 2001; Galley et al., 2004); in
 10224 **tree-to-string translation**, the source side is parsed, and then transformed into a string on
 10225 the target side (Liu et al., 2006). A key question for syntax-based translation is the extent
 10226 to which we phrasal constituents align across translations (Fox, 2002), because this gov-
 10227 erns the extent to which we can rely on monolingual parsers and treebanks. For more on
 10228 syntax-based machine translation, see the monograph by Williams et al. (2016).

10229 18.3 Neural machine translation

Neural network models for machine translation are based on the **encoder-decoder** architecture (Cho et al., 2014). The encoder network converts the source language sentence into a vector or matrix representation; the decoder network then converts the encoding into a sentence in the target language.

$$\mathbf{z} = \text{ENCODE}(\mathbf{w}^{(s)}) \quad [18.24]$$

$$\mathbf{w}^{(t)} \mid \mathbf{w}^{(s)} \sim \text{DECODE}(\mathbf{z}), \quad [18.25]$$

10230 where the second line means that the function $\text{DECODE}(\mathbf{z})$ defines the conditional proba-
 10231 bility $p(\mathbf{w}^{(t)} \mid \mathbf{w}^{(s)})$.

The decoder is typically a recurrent neural network, which generates the target language sentence one word at a time, while recurrently updating a hidden state. The encoder and decoder networks are trained end-to-end from parallel sentences. If the output layer of the decoder is a logistic function, then the entire architecture can be trained to maximize the conditional log-likelihood,

$$\log p(\mathbf{w}^{(t)} \mid \mathbf{w}^{(s)}) = \sum_{m=1}^{M^{(t)}} p(w_m^{(t)} \mid \mathbf{w}_{1:m-1}^{(t)}, \mathbf{z}) \quad [18.26]$$

$$p(w_m^{(t)} \mid \mathbf{w}_{1:m-1}^{(t)}, \mathbf{w}^{(s)}) \propto \exp\left(\boldsymbol{\beta}_{w_m^{(t)}} \cdot \mathbf{h}_{m-1}^{(t)}\right) \quad [18.27]$$

where the hidden state $\mathbf{h}_{m-1}^{(t)}$ is a recurrent function of the previously generated text $\mathbf{w}_{1:m-1}^{(t)}$ and the encoding \mathbf{z} . The second line is equivalent to writing,

$$w_m^{(t)} \mid \mathbf{w}_{1:m-1}^{(t)}, \mathbf{w}^{(s)} \sim \text{SoftMax}\left(\boldsymbol{\beta} \cdot \mathbf{h}_{m-1}^{(t)}\right), \quad [18.28]$$

10232 where $\boldsymbol{\beta} \in \mathbb{R}^{(V^{(t)} \times K)}$ is the matrix of output word vectors for the $V^{(t)}$ words in the target
 10233 language vocabulary.

The simplest encoder-decoder architecture is the **sequence-to-sequence model** (Sutskever et al., 2014). In this model, the encoder is set to the final hidden state of a **long short-term**

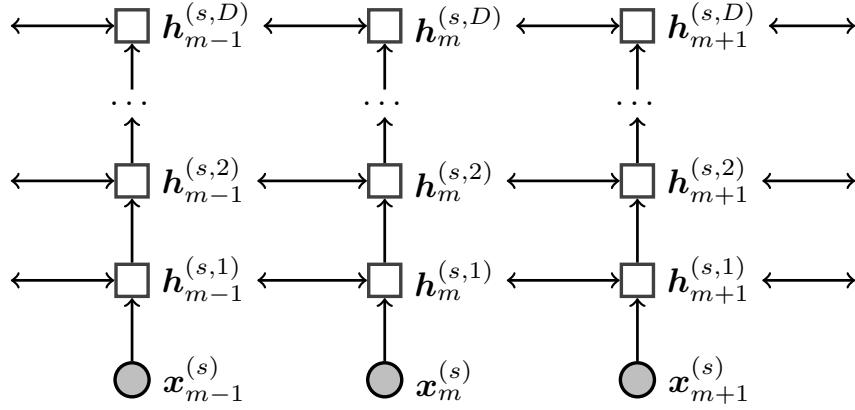


Figure 18.5: A deep bidirectional LSTM encoder

memory (LSTM) (see § 6.3.3) on the source sentence:

$$\mathbf{h}_m^{(s)} = \text{LSTM}(\mathbf{x}_m^{(s)}, \mathbf{h}_{m-1}^{(s)}) \quad [18.29]$$

$$\mathbf{z} \triangleq \mathbf{h}_{M^{(s)}}^{(s)}, \quad [18.30]$$

where $\mathbf{x}_m^{(s)}$ is the embedding of source language word $w_m^{(s)}$. The encoding then provides the initial hidden state for the decoder LSTM:

$$\mathbf{h}_0^{(t)} = \mathbf{z} \quad [18.31]$$

$$\mathbf{h}_m^{(t)} = \text{LSTM}(\mathbf{x}_m^{(t)}, \mathbf{h}_{m-1}^{(t)}), \quad [18.32]$$

10234 where $\mathbf{x}_m^{(t)}$ is the embedding of the target language word $w_m^{(t)}$.

10235 Sequence-to-sequence translation is nothing more than wiring together two LSTMs:
 10236 one to read the source, and another to generate the target. To make the model work well,
 10237 some additional tweaks are needed:

- 10238 • Most notably, the model works much better if the source sentence is reversed, reading
 10239 from the end of the sentence back to the beginning. In this way, the words at the
 10240 beginning of the source have the greatest impact on the encoding \mathbf{z} , and therefore
 10241 impact the words at the beginning of the target sentence. Later work on more advanced
 10242 encoding models, such as **neural attention** (see § 18.3.1), has eliminated the
 10243 need for reversing the source sentence.
- The encoder and decoder can be implemented as **deep LSTMs**, with multiple layers of hidden states. As shown in Figure 18.5, each hidden state $\mathbf{h}_m^{(s,i)}$ at layer i is treated

as the input to an LSTM at layer $i + 1$:

$$\mathbf{h}_m^{(s,1)} = \text{LSTM}(\mathbf{x}_m^{(s)}, \mathbf{h}_{m-1}^{(s)}) \quad [18.33]$$

$$\mathbf{h}_m^{(s,i+1)} = \text{LSTM}(\mathbf{h}_m^{(s,i)}, \mathbf{h}_{m-1}^{(s,i+1)}), \quad \forall i \geq 1. \quad [18.34]$$

10244 The original work on sequence-to-sequence translation used four layers; in 2016,
 10245 Google’s commercial machine translation system used eight layers (Wu et al., 2016).⁴

- 10246 • Significant improvements can be obtained by creating an **ensemble** of translation
 10247 models, each trained from a different random initialization. For an ensemble of size
 10248 N , the per-token decoding probability is set equal to,

$$p(w^{(t)} | z, \mathbf{w}_{1:m-1}^{(t)}) = \frac{1}{N} \sum_{i=1}^N p_i(w^{(t)} | z, \mathbf{w}_{1:m-1}^{(t)}), \quad [18.35]$$

10249 where p_i is the decoding probability for model i . Each translation model in the
 10250 ensemble includes its own encoder and decoder networks.

- 10251 • The original sequence-to-sequence model used a fairly standard training setup: stochastic
 10252 gradient descent with an exponentially decreasing learning rate after the first five
 10253 epochs; mini-batches of 128 sentences, chosen to have similar length so that each
 10254 sentence on the batch will take roughly the same amount of time to process; gradient
 10255 clipping (see § 3.3.4) to ensure that the norm of the gradient never exceeds some
 10256 predefined value.

10257 18.3.1 Neural attention

10258 The sequence-to-sequence model discussed in the previous section was a radical departure
 10259 from statistical machine translation, in which each word or phrase in the target lan-
 10260 guage is conditioned on a single word or phrase in the source language. Both approaches
 10261 have advantages. Statistical translation leverages the idea of compositionality — transla-
 10262 tions of large units should be based on the translations of their component parts — and
 10263 this seems crucial if we are to scale translation to longer units of text. But the translation
 10264 of each word or phrase often depends on the larger context, and encoder-decoder models
 10265 capture this context at the sentence level.

10266 Is it possible for translation to be both contextualized and compositional? One ap-
 10267 proach is to augment neural translation with an **attention mechanism**. The idea of neural
 10268 attention was described in § 17.5, but its application to translation bears further discus-
 10269 sion. In general, attention can be thought of as using a query to select from a memory
 10270 of key-value pairs. However, the query, keys, and values are all vectors, and the entire

⁴Google reports that this system took six days to train for English-French translation, using 96 NVIDIA K80 GPUs, which would have cost roughly half a million dollars at the time.

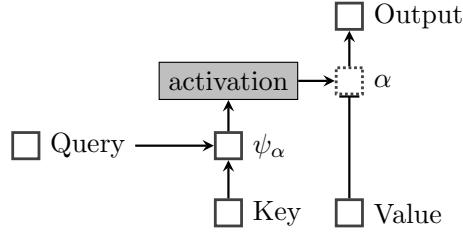


Figure 18.6: A general view of neural attention. The dotted box indicates that each $\alpha_{m \rightarrow n}$ can be viewed as a **gate** on value n .

operation is differentiable. For each key n in the memory, we compute a score $\psi_\alpha(m, n)$ with respect to the query m . That score is a function of the compatibility of the key and the query, and can be computed using a small feedforward neural network. The vector of scores is passed through an activation function, such as softmax. The output of this activation function is a vector of non-negative numbers $[\alpha_{m \rightarrow 1}, \alpha_{m \rightarrow 2}, \dots, \alpha_{m \rightarrow N}]^\top$, with length N equal to the size of the memory. Each value in the memory v_n is multiplied by the attention $\alpha_{m \rightarrow n}$; the sum of these scaled values is the output. This process is shown in Figure 18.6. In the extreme case that $\alpha_{m \rightarrow n} = 1$ and $\alpha_{m \rightarrow n'} = 0$ for all other n' , then the attention mechanism simply selects the value v_n from the memory.

Neural attention makes it possible to integrate alignment into the encoder-decoder architecture. Rather than encoding the entire source sentence into a fixed length vector z , it can be encoded into a matrix $Z \in \mathbb{R}^{K \times M^{(S)}}$, where K is the dimension of the hidden state, and $M^{(S)}$ is the number of tokens in the source input. Each column of Z represents the state of a recurrent neural network over the source sentence. These vectors are constructed from a **bidirectional LSTM** (see § 7.6), which can be a deep network as shown in Figure 18.5. These columns are both the keys and the values in the attention mechanism.

At each step m in decoding, the attentional state is computed by executing a query, which is equal to the state of the decoder, $h_m^{(t)}$. The resulting compatibility scores are,

$$\psi_\alpha(m, n) = v_\alpha \cdot \tanh(\Theta_\alpha[h_m^{(t)}; h_n^{(s)}]). \quad [18.36]$$

The function ψ is thus a two layer feedforward neural network, with weights v_α on the output layer, and weights Θ_α on the input layer. To convert these scores into attention weights, we apply an activation function, which can be vector-wise softmax or an element-wise sigmoid:

Softmax attention

$$\alpha_{m \rightarrow n} = \frac{\exp \psi_\alpha(m, n)}{\sum_{n'=1}^{M^{(s)}} \exp \psi_\alpha(m, n')} \quad [18.37]$$

Sigmoid attention

$$\alpha_{m \rightarrow n} = \sigma(\psi_\alpha(m, n)) \quad [18.38]$$

The attention α is then used to compute a context vector c_m by taking a weighted average over the columns of Z ,

$$c_m = \sum_{n=1}^{M^{(s)}} \alpha_{m \rightarrow n} z_n, \quad [18.39]$$

where $\alpha_{m \rightarrow n} \in [0, 1]$ is the amount of attention from word m of the target to word n of the source. The context vector can be incorporated into the decoder’s word output probability model, by adding another layer to the decoder (Luong et al., 2015):

$$\tilde{h}_m^{(t)} = \tanh(\Theta_c[h_m^{(t)}; c_m]) \quad [18.40]$$

$$p(w_{m+1}^{(t)} | w_{1:m}^{(t)}, w^{(s)}) \propto \exp\left(\beta_{w_{m+1}^{(t)}} \cdot \tilde{h}_m^{(t)}\right). \quad [18.41]$$

10291 Here the decoder state $h_m^{(t)}$ is concatenated with the context vector, forming the input
 10292 to compute a final output vector $\tilde{h}_m^{(t)}$. The context vector can be incorporated into the
 10293 decoder recurrence in a similar manner (Bahdanau et al., 2014).

10294 **18.3.2 *Neural machine translation without recurrence**

In the encoder-decoder model, attention’s “keys and values” are the hidden state representations in the encoder network, z , and the “queries” are state representations in the decoder network $h^{(t)}$. It is also possible to completely eliminate recurrence from neural translation, by applying **self-attention** (Lin et al., 2017; Kim et al., 2017) within the encoder and decoder, as in the **transformer architecture** (Vaswani et al., 2017). For level i , the basic equations of the encoder side of the transformer are:

$$z_m^{(i)} = \sum_{n=1}^{M^{(s)}} \alpha_{m \rightarrow n}^{(i)} (\Theta_v h_n^{(i-1)}) \quad [18.42]$$

$$h_m^{(i)} = \Theta_2 \text{ReLU}(\Theta_1 z_m^{(i)} + b_1) + b_2. \quad [18.43]$$

10295 For each token m at level i , we compute self-attention over the entire source sentence.
 10296 The keys, values, and queries are all projections of the vector $h^{(i-1)}$: for example, in Equa-
 10297 tion 18.42, the value v_n is the projection $\Theta_v h_n^{(i-1)}$. The attention scores $\alpha_{m \rightarrow n}^{(i)}$ are com-
 10298 puted using a scaled form of softmax attention,

$$\alpha_{m \rightarrow n} \propto \exp(\psi_\alpha(m, n)/M), \quad [18.44]$$

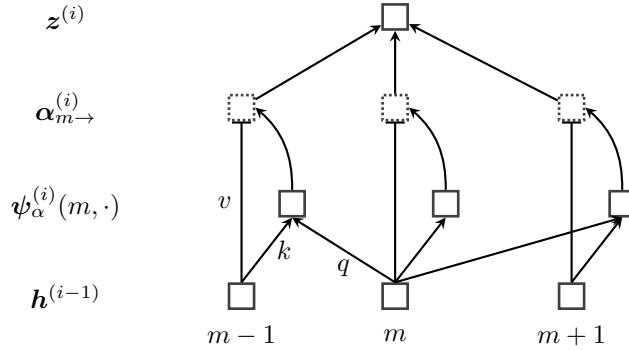


Figure 18.7: The transformer encoder’s computation of $z_m^{(i)}$ from $\mathbf{h}^{(i-1)}$. The key, value, and query are shown for token $m - 1$. For example, $\psi_\alpha^{(i)}(m, m - 1)$ is computed from the key $\Theta_k \mathbf{h}_{m-1}^{(i-1)}$ and the query $\Theta_q \mathbf{h}_m^{(i-1)}$, and the gate $\alpha_{m \rightarrow m-1}^{(i)}$ operates on the value $\Theta_v \mathbf{h}_{m-1}^{(i-1)}$. The figure shows a minimal version of the architecture, with a single attention head. With multiple heads, it is possible to attend to different properties of multiple words.

where M is the length of the input. This encourages the attention to be more evenly dispersed across the input. Self-attention is applied across multiple “heads”, each using different projections of $\mathbf{h}^{(i-1)}$ to form the keys, values, and queries. This architecture is shown in Figure 18.7. The output of the self-attentional layer is the representation $z_m^{(i)}$, which is then passed through a two-layer feed-forward network, yielding the input to the next layer, $\mathbf{h}^{(i)}$. This self-attentional architecture can be applied in the decoder as well, but this requires that there is zero attention to future words: $\alpha_{m \rightarrow n} = 0$ for all $n > m$.

To ensure that information about word order in the source is integrated into the model, the encoder augments the base layer of the network with **positional encodings** of the indices of each word in the source. These encodings are vectors for each position $m \in \{1, 2, \dots, M\}$. The transformer sets these encodings equal to a set of sinusoidal functions of m ,

$$e_{2i-1}(m) = \sin(m/(10000^{\frac{2i}{K_e}})) \quad [18.45]$$

$$e_{2i}(m) = \cos(m/(10000^{\frac{2i}{K_e}})), \quad \forall i \in \{1, 2, \dots, K_e/2\} \quad [18.46]$$

where $e_{2i}(m)$ is the value at element $2i$ of the encoding for index m . As we progress through the encoding, the sinusoidal functions have progressively narrower bandwidths. This enables the model to learn to attend by relative positions of words. The positional encodings are concatenated with the word embeddings x_m at the base layer of the model.⁵

⁵The transformer architecture relies on several additional tricks, including **layer normalization** (see

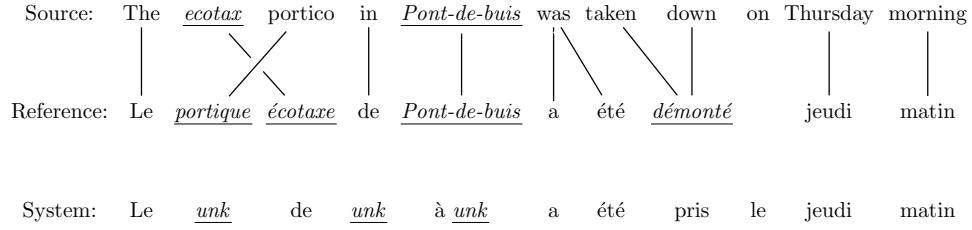


Figure 18.8: Translation with *unknown words*. The system outputs *unk* to indicate words that are outside its vocabulary. Figure adapted from Luong et al. (2015).

10310 **Convolutional neural networks** (see § 3.4) have also been applied as encoders in neu-
 10311 ral machine translation. For each word $w_m^{(s)}$, a convolutional network computes a rep-
 10312 resentation $h_m^{(s)}$ from the embeddings of the word and its neighbors. This procedure is
 10313 applied several times, creating a deep convolutional network. The recurrent decoder then
 10314 computes a set of attention weights over these convolutional representations, using the
 10315 decoder’s hidden state $h^{(t)}$ as the queries. This attention vector is used to compute a
 10316 weighted average over the outputs of *another* convolutional neural network of the source,
 10317 yielding an averaged representation c_m , which is then fed into the decoder. As with the
 10318 transformer, speed is the main advantage over recurrent encoding models; another sim-
 10319 ilarity is that word order information is approximated through the use of positional en-
 10320 codings.⁶

10321 18.3.3 Out-of-vocabulary words

10322 Thus far, we have treated translation as a problem at the level of words or phrases. For
 10323 words that do not appear in the training data, all such models will struggle. There are
 10324 two main reasons for the presence of out-of-vocabulary (OOV) words:

- 10325 • New proper nouns, such as family names or organizations, are constantly arising —
 particularly in the news domain. The same is true, to a lesser extent, for technical
 terminology. This issue is shown in Figure 18.8.
- 10328 • In many languages, words have complex internal structure, known as **morphology**.
 An example is German, which uses compounding to form nouns like *Abwasserbe-*
 handlungsanlage (*sewage water treatment plant*; example from Sennrich et al. (2016)).

§ 3.3.4), residual connections around the nonlinear activations (see § 3.2.2), and a non-monotonic learning rate schedule.

⁶A recent evaluation found that best performance was obtained by using a recurrent network for the decoder, and a transformer for the encoder (Chen et al., 2018). The transformer was also found to significantly outperform a convolutional neural network.

10331 While compounds could in principle be addressed by better tokenization (see § 8.4),
 10332 other morphological processes involve more complex transformations of subword
 10333 units.

10334 Names and technical terms can be handled in a postprocessing step: after first identi-
 10335 fying alignments between unknown words in the source and target, we can look up each
 10336 aligned source word in a dictionary, and choose a replacement (Luong et al., 2015). If the
 10337 word does not appear in the dictionary, it is likely to be a proper noun, and can be copied
 10338 directly from the source to the target. This approach can also be integrated directly into
 10339 the translation model, rather than applying it as a postprocessing step (Jean et al., 2015).

10340 Words with complex internal structure can be handled by translating subword units
 10341 rather than entire words. A popular technique for identifying subword units is **byte-pair**
 10342 **encoding** (BPE; Gage, 1994; Sennrich et al., 2016). The initial vocabulary is defined as the
 10343 set of characters used in the text. The most common character bigram is then merged into
 10344 a new symbol, the vocabulary is updated, and the merging operation is applied again. For
 10345 example, given the dictionary $\{fish, fished, want, wanted, bike, biked\}$, we would first form
 10346 the subword unit ed , since this character bigram appears in three of the six words. Next,
 10347 there are several bigrams that each appear in a pair of words: fi, is, sh, wa, an , etc. These can
 10348 be merged in any order. By iterating this process, we eventually reach the segmentation,
 10349 $\{fish, fish+ed, want, want+ed, bik+e, bik+ed\}$. At this point, there are no bigrams that appear
 10350 more than once. In real data, merging is performed until the number of subword units
 10351 reaches some predefined threshold, such as 10^4 .

10352 Each subword unit is treated as a token for translation, in both the encoder (source
 10353 side) and decoder (target side). BPE can be applied jointly to the union of the source and
 10354 target vocabularies, identifying subword units that appear in both languages. For lan-
 10355 guages that have different scripts, such as English and Russian, **transliteration** between
 10356 the scripts should be applied first.⁷

10357 18.4 Decoding

Given a trained translation model, the decoding task is:

$$\hat{\mathbf{w}}^{(t)} = \underset{\mathbf{w} \in \mathcal{V}^*}{\operatorname{argmax}} \Psi(\mathbf{w}, \mathbf{w}^{(s)}), \quad [18.47]$$

10358 where $\mathbf{w}^{(t)}$ is a sequence of tokens from the target vocabulary \mathcal{V} . It is not possible to
 10359 efficiently obtain exact solutions to the decoding problem, for even minimally effective

⁷Transliteration is crucial for converting names and other foreign words between languages that do not share a single script, such as English and Japanese. It is typically approached using the finite-state methods discussed in chapter 9 (Knight and Graehl, 1998).

models in either statistical or neural machine translation. Today’s state-of-the-art translation systems use **beam search** (see § 11.3.1), which is an incremental decoding algorithm that maintains a small constant number of competitive hypotheses. Such greedy approximations are reasonably effective in practice, and this may be in part because the decoding objective is only loosely correlated with measures of translation quality, so that exact optimization of [18.47] may not greatly improve the resulting translations.

Decoding in neural machine translation is simpler than in phrase-based statistical machine translation.⁸ The scoring function Ψ is defined,

$$\Psi(\mathbf{w}^{(t)}, \mathbf{w}^{(s)}) = \sum_{m=1}^{M^{(t)}} \psi(w_m^{(t)}; \mathbf{w}_{1:m-1}^{(t)}, \mathbf{z}) \quad [18.48]$$

$$\psi(w^{(t)}; \mathbf{w}_{1:m-1}^{(t)}, \mathbf{z}) = \beta_{w_m^{(t)}} \cdot \mathbf{h}_m^{(t)} - \log \sum_{w \in \mathcal{V}} \exp(\beta_w \cdot \mathbf{h}_m^{(t)}), \quad [18.49]$$

where \mathbf{z} is the encoding of the source sentence $\mathbf{w}^{(s)}$, and $\mathbf{h}_m^{(t)}$ is a function of the encoding \mathbf{z} and the decoding history $\mathbf{w}_{1:m-1}^{(t)}$. This formulation subsumes the attentional translation model, where \mathbf{z} is a matrix encoding of the source.

Now consider the incremental decoding algorithm,

$$\hat{w}_m^{(t)} = \operatorname{argmax}_{w \in \mathcal{V}} \psi(w; \hat{\mathbf{w}}_{1:m-1}^{(t)}, \mathbf{z}), \quad m = 1, 2, \dots \quad [18.50]$$

This algorithm selects the best target language word at position m , assuming that it has already generated the sequence $\hat{\mathbf{w}}_{1:m-1}^{(t)}$. (Termination can be handled by augmenting the vocabulary \mathcal{V} with a special end-of-sequence token, ■.) The incremental algorithm is likely to produce a suboptimal solution to the optimization problem defined in Equation 18.47, because selecting the highest-scoring word at position m can set the decoder on a “garden path,” in which there are no good choices at some later position $n > m$. We might hope for some dynamic programming solution, as in sequence labeling (§ 7.3). But the Viterbi algorithm and its relatives rely on a Markov decomposition of the objective function into a sum of local scores: for example, scores can consider locally adjacent tags (y_m, y_{m-1}) , but not the entire tagging history $\mathbf{y}_{1:m}$. This decomposition is not applicable to recurrent neural networks, because the hidden state $\mathbf{h}_m^{(t)}$ is impacted by the entire history $\mathbf{w}_{1:m}^{(t)}$; this sensitivity to long-range context is precisely what makes recurrent neural networks so effective.⁹ In fact, it can be shown that decoding from any recurrent neural network is NP-complete (Siegelmann and Sontag, 1995; Chen et al., 2018).

⁸For more on decoding in phrase-based statistical models, see Koehn (2009).

⁹Note that this problem does not impact RNN-based sequence labeling models (see § 7.6). This is because the tags produced by these models do not affect the recurrent state.

10383 **Beam search** Beam search is a general technique for avoiding search errors when ex-
 10384 haustive search is impossible; it was first discussed in § 11.3.1. Beam search can be seen
 10385 as a variant of the incremental decoding algorithm sketched in Equation 18.50, but at
 10386 each step m , a set of K different hypotheses are kept on the beam. For each hypothesis
 10387 $k \in \{1, 2, \dots, K\}$, we compute both the current score $\sum_{m=1}^{M(t)} \psi(w_{k,m}^{(t)}; \mathbf{w}_{k,1:m-1}^{(t)}, \mathbf{z})$ as well as
 10388 the current hidden state $\mathbf{h}_k^{(t)}$. At each step in the beam search, the K top-scoring children
 10389 of each hypothesis currently on the beam are “expanded”, and the beam is updated. For
 10390 a detailed description of beam search for RNN decoding, see Graves (2012).

10391 **Learning and search** Conventionally, the learning algorithm is trained to predict the right token in the translation, conditioned on the translation history being correct. But if decoding must be approximate, then we might do better by modifying the learning algorithm to be robust to errors in the translation history. **Scheduled sampling** does this by training on histories that sometimes come from the ground truth, and sometimes come from the model’s own output (Bengio et al., 2015).¹⁰ As training proceeds, the training wheels come off: we increase the fraction of tokens that come from the model rather than the ground truth. Another approach is to train on an objective that relates directly to beam search performance (Wiseman et al., 2016). **Reinforcement learning** has also been applied to decoding of RNN-based translation models, making it possible to directly optimize translation metrics such as BLEU (Ranzato et al., 2016).

10402 18.5 Training towards the evaluation metric

10403 In likelihood-based training, the objective is the maximize the probability of a parallel corpus. However, translations are not evaluated in terms of likelihood: metrics like BLEU consider only the correctness of a single output translation, and not the range of probabilities that the model assigns. It might therefore be better to train translation models to achieve the highest BLEU score possible — to the extent that we believe BLEU measures translation quality. Unfortunately, BLEU and related metrics are not friendly for optimization: they are discontinuous, non-differentiable functions of the parameters of the translation model.

Consider an error function $\Delta(\hat{\mathbf{w}}^{(t)}, \mathbf{w}^{(t)})$, which measures the discrepancy between the system translation $\hat{\mathbf{w}}^{(t)}$ and the reference translation $\mathbf{w}^{(t)}$; this function could be based on BLEU or any other metric on translation quality. One possible criterion would be to select

¹⁰Scheduled sampling builds on earlier work on learning to search (Daumé III et al., 2009; Ross et al., 2011), which are also described in § 15.2.4.

the parameters θ that minimize the error of the system’s preferred translation,

$$\hat{\mathbf{w}}^{(t)} = \operatorname{argmax}_{\mathbf{w}^{(t)}} \Psi(\mathbf{w}^{(t)}, \mathbf{w}^{(s)}; \theta) \quad [18.51]$$

$$\hat{\theta} = \operatorname{argmin}_{\theta} \Delta(\hat{\mathbf{w}}^{(t)}, \mathbf{w}^{(s)}) \quad [18.52]$$

10411 However, identifying the top-scoring translation $\hat{\mathbf{w}}^{(t)}$ is usually intractable, as described
 10412 in the previous section. In **minimum error-rate training (MERT)**, $\hat{\mathbf{w}}^{(t)}$ is selected from a
 10413 set of candidate translations $\mathcal{Y}(\mathbf{w}^{(s)})$; this is typically a strict subset of all possible transla-
 10414 tions, so that it is only possible to optimize an approximation to the true error rate (Och
 10415 and Ney, 2003).

A further issue is that the objective function in Equation 18.52 is discontinuous and non-differentiable, due to the argmax over translations: an infinitesimal change in the parameters θ could cause another translation to be selected, with a completely different error. To address this issue, we can instead minimize the **risk**, which is defined as the expected error rate,

$$R(\theta) = E_{\hat{\mathbf{w}}^{(t)} | \mathbf{w}^{(s)}; \theta} [\Delta(\hat{\mathbf{w}}^{(t)}, \mathbf{w}^{(t)})] \quad [18.53]$$

$$= \sum_{\hat{\mathbf{w}}^{(t)} \in \mathcal{Y}(\mathbf{w}^{(s)})} p(\hat{\mathbf{w}}^{(t)} | \mathbf{w}^{(s)}) \times \Delta(\hat{\mathbf{w}}^{(t)}, \mathbf{w}^{(t)}). \quad [18.54]$$

10416 **Minimum risk training** minimizes the sum of $R(\theta)$ across all instances in the training set.

The risk can be generalized by exponentiating the translation probabilities,

$$\tilde{p}(\mathbf{w}^{(t)}; \theta, \alpha) \propto \left(p(\mathbf{w}^{(t)} | \mathbf{w}^{(s)}; \theta) \right)^\alpha \quad [18.55]$$

$$\tilde{R}(\theta) = \sum_{\hat{\mathbf{w}}^{(t)} \in \mathcal{Y}(\mathbf{w}^{(s)})} \tilde{p}(\hat{\mathbf{w}}^{(t)} | \mathbf{w}^{(s)}; \alpha, \theta) \times \Delta(\hat{\mathbf{w}}^{(t)}, \mathbf{w}^{(t)}) \quad [18.56]$$

10417 where $\mathcal{Y}(\mathbf{w}^{(s)})$ is now the set of *all* possible translations for $\mathbf{w}^{(s)}$. Exponentiating the prob-
 10418 abilities in this way is known as **annealing** (Smith and Eisner, 2006). When $\alpha = 1$, then
 10419 $\tilde{R}(\theta) = R(\theta)$; when $\alpha = \infty$, then $\tilde{R}(\theta)$ is equivalent to the sum of the errors of the maxi-
 10420 mum probability translations for each sentence in the dataset.

Clearly the set of candidate translations $\mathcal{Y}(\mathbf{w}^{(s)})$ is too large to explicitly sum over. Because the error function Δ generally does not decompose into smaller parts, there is no efficient dynamic programming solution to sum over this set. We can approximate the sum $\sum_{\hat{\mathbf{w}}^{(t)} \in \mathcal{Y}(\mathbf{w}^{(s)})}$ with a sum over a finite number of samples, $\{\mathbf{w}_1^{(t)}, \mathbf{w}_2^{(t)}, \dots, \mathbf{w}_K^{(t)}\}$. If these samples were drawn uniformly at random, then the (annealed) risk would be

approximated as (Shen et al., 2016),

$$\tilde{R}(\boldsymbol{\theta}) \approx \frac{1}{Z} \sum_{k=1}^K \tilde{p}(\mathbf{w}_k^{(t)} | \mathbf{w}^{(s)}; \boldsymbol{\theta}, \alpha) \times \Delta(\mathbf{w}_k^{(t)}, \mathbf{w}^{(t)}) \quad [18.57]$$

$$Z = \sum_{k=1}^K \tilde{p}(\mathbf{w}_k^{(t)} | \mathbf{w}^{(s)}; \boldsymbol{\theta}, \alpha). \quad [18.58]$$

10421 Shen et al. (2016) report that performance plateaus at $K = 100$ for minimum risk training
 10422 of neural machine translation.

Uniform sampling over the set of all possible translations is undesirable, because most translations have very low probability. A solution from Monte Carlo estimation is **importance sampling**, in which we draw samples from a **proposal distribution** $q(\mathbf{w}^{(s)})$. This distribution can be set equal to the current translation model $p(\mathbf{w}^{(t)} | \mathbf{w}^{(s)}; \boldsymbol{\theta})$. Each sample is then weighted by an **importance score**, $\omega_k = \frac{\tilde{p}(\mathbf{w}_k^{(t)} | \mathbf{w}^{(s)})}{q(\mathbf{w}_k^{(t)}; \mathbf{w}^{(s)})}$. The effect of this weighting is to correct for any mismatch between the proposal distribution q and the true distribution \tilde{p} . The risk can then be approximated as,

$$\mathbf{w}_k^{(t)} \sim q(\mathbf{w}^{(s)}) \quad [18.59]$$

$$\omega_k = \frac{\tilde{p}(\mathbf{w}_k^{(t)} | \mathbf{w}^{(s)})}{q(\mathbf{w}_k^{(t)}; \mathbf{w}^{(s)})} \quad [18.60]$$

$$\tilde{R}(\boldsymbol{\theta}) \approx \frac{1}{\sum_{k=1}^K \omega_k} \sum_{k=1}^K \omega_k \times \Delta(\mathbf{w}_k^{(t)}, \mathbf{w}^{(t)}). \quad [18.61]$$

10423 Importance sampling will generally give a more accurate approximation than uniform
 10424 sampling. The only formal requirement is that the proposal assigns non-zero probability
 10425 to every $\mathbf{w}^{(t)} \in \mathcal{Y}(\mathbf{w}^{(s)})$. For more on importance sampling and related methods, see
 10426 Robert and Casella (2013).

10427 Additional resources

10428 A complete textbook on machine translation is available from Koehn (2009). While this
 10429 book precedes recent work on neural translation, a more recent draft chapter on neural
 10430 translation models is also available (Koehn, 2017). Neubig (2017) provides a com-
 10431 prehensive tutorial on neural machine translation, starting from first principles. The course
 10432 notes from Cho (2015) are also useful. Several neural machine translation libraries are
 10433 available: LAMTRAM is an implementation of neural machine translation in DYNET (Neu-
 10434 big et al., 2017); OPENNMT (Klein et al., 2017) and FAIRSEQ are available in PYTORCH;

10435 TENSOR2TENSOR is an implementation of several of the Google translation models in TEN-
 10436 SORFLOW (Abadi et al., 2016).

10437 Literary translation is especially challenging, even for expert human translators. Mes-
 10438 sud (2014) describes some of these issues in her review of an English translation of *L'étranger*,
 10439 the 1942 French novel by Albert Camus.¹¹ She compares the new translation by Sandra
 10440 Smith against earlier translations by Stuart Gilbert and Matthew Ward, focusing on the
 10441 difficulties presented by a single word in the first sentence:

10442 Then, too, Smith has reconsidered the book's famous opening. Camus's
 10443 original is deceptively simple: "*Aujourd'hui, maman est morte.*" Gilbert influ-
 10444 enced generations by offering us "Mother died today"—inscribing in Meur-
 10445 sault [the narrator] from the outset a formality that could be construed as
 10446 heartlessness. But *maman*, after all, is intimate and affectionate, a child's name
 10447 for his mother. Matthew Ward concluded that it was essentially untranslatable
 10448 ("mom" or "mummy" being not quite apt), and left it in the original French:
 10449 "Maman died today." There is a clear logic in this choice; but as Smith has
 10450 explained, in an interview in *The Guardian*, *maman* "didn't really tell the reader
 10451 anything about the connotation." She, instead, has translated the sentence as
 10452 "My mother died today."

10453 I chose "My mother" because I thought about how someone would
 10454 tell another person that his mother had died. Meursault is speaking
 10455 to the reader directly. "My mother died today" seemed to me the
 10456 way it would work, and also implied the closeness of "maman" you
 10457 get in the French.

10458 Elsewhere in the book, she has translated *maman* as "mama" — again, striving
 10459 to come as close as possible to an actual, colloquial word that will carry the
 10460 same connotations as *maman* does in French.

10461 The passage is a reminder that while the quality of machine translation has improved
 10462 dramatically in recent years, expert human translations draw on considerations that are
 10463 beyond the ken of any contemporary computational approach.

10464 Exercises

- 10465 1. Using Google translate or another online service, translate the following example
 10466 into two different languages of your choice:

¹¹The book review is currently available online at <http://www.nybooks.com/articles/2014/06/05/camus-new-letranger/>.

10467 (18.4) It is not down on any map; true places never are.

10468 Then translate each result back into English. Which is closer to the original? Can
10469 you explain the differences?

- 10470 2. Compute the unsmoothed n -gram precisions $p_1 \dots p_4$ for the two back-translations
10471 in the previous problem, using the original source as the reference. Your n -grams
10472 should include punctuation, and you should segment conjunctions like *it's* into two
10473 tokens.
- 10474 3. You are given the following dataset of translations from “simple” to “difficult” En-
10475 glish:

- 10476 (18.5) a. *Kids like cats.*
Children adore felines.
10477 b. *Cats hats.*
Felines fedoras.

10478 Estimate a word-to-word statistical translation model from simple English (source)
10479 to difficult English (target), using the expectation-maximization as described in § 18.2.2.
10480 Compute two iterations of the algorithm by hand, starting from a uniform transla-
10481 tion model, and using the simple alignment model $p(a_m | m, M^{(s)}, M^{(t)}) = \frac{1}{M^{(t)}}$.
10482 Hint: in the final M-step, you will want to switch from fractions to decimals.

- 10483 4. Building on the previous problem, what will be the converged translation proba-
10484 bility table? Can you state a general condition about the data, under which this
10485 translation model will fail in the way that it fails here?
- 10486 5. Propose a simple alignment model that would make it possible to recover the correct
10487 translation probabilities from the toy dataset in the previous two problems.
- 10488 6. Let $\ell_{m+1}^{(t)}$ represent the loss at word $m+1$ of the target, and let $h_n^{(s)}$ represent the hid-
10489 den state at word n of the source. Write the expression for the derivative $\frac{\partial \ell_{m+1}^{(t)}}{\partial h_n^{(s)}}$ in the
10490 sequence-to-sequence translation model expressed in Equations [18.29-18.32]. You
10491 may assume that both the encoder and decoder are one-layer LSTMs. In general,
10492 how many terms are on the shortest backpropagation path from $\ell_{m+1}^{(t)}$ to $h_n^{(s)}$?
- 10493 7. Now consider the neural attentional model from § 18.3.1, with sigmoid attention.
10494 The derivative $\frac{\partial \ell_{m+1}^{(t)}}{\partial z_n}$ is the sum of many paths through the computation graph;
10495 identify the shortest such path. You may assume that the initial state of the decoder
10496 recurrence $h_0^{(t)}$ is *not* tied to the final state of the encoder recurrence $h_{M^{(s)}}^{(s)}$.

- 10497 8. Apply byte-pair encoding for the vocabulary *it*, *unit*, *unite*, until no bigram appears
 10498 more than once.
- 10499 9. This problem relates to the complexity of machine translation. Suppose you have
 10500 an oracle that returns the list of words to include in the translation, so that your
 10501 only task is to order the words. Furthermore, suppose that the scoring function
 10502 over orderings is a sum over bigrams, $\sum_{m=1}^M \psi(\mathbf{w}_m^{(t)}, \mathbf{w}_{m-1}^{(t)})$. Show that the problem
 10503 of finding the optimal translation is NP-complete, by reduction from a well-known
 10504 problem.
- 10505 10. Hand-design an attentional recurrent translation model that simply copies the input
 10506 from the source to the target. You may assume an arbitrarily large hidden state, and
 10507 you may assume that there is a finite maximum input length M . Specify all the
 10508 weights such that the maximum probability translation of any source is the source
 10509 itself. Hint: it is simplest to use the Elman recurrence $\mathbf{h}_m = f(\Theta \mathbf{h}_{m-1} + \mathbf{x}_m)$ rather
 10510 than an LSTM.
- 10511 11. Give a synchronized derivation (§ 18.2.4) for the Spanish-English translation,

10512 (18.6) *El pez enojado atacado.*
 The fish angry attacked.
 10513 The angry fish attacked.

10514 As above, the second line shows a word-for-word gloss, and the third line shows
 10515 the desired translation. Use the synchronized production rule in [18.22], and design
 10516 the other production rules necessary to derive this sentence pair. You may derive
 10517 (*atacado*, *attacked*) directly from VP.

10518

Chapter 19

10519

Text generation

10520 In many of the most interesting problems in natural language processing, language is
10521 the output. The previous chapter described the specific case of machine translation, but
10522 there are many other applications, from summarization of research articles, to automated
10523 journalism, to dialogue systems. This chapter emphasizes three main scenarios: data-to-
10524 text, in which text is generated to explain or describe a structured record or unstructured
10525 perceptual input; text-to-text, which typically involves fusing information from multiple
10526 linguistic sources into a single coherent summary; and dialogue, in which text is generated
10527 as part of an interactive conversation with one or more human participants.

10528

19.1 Data-to-text generation

10529 In data-to-text generation, the input ranges from structured records, such as the descrip-
10530 tion of an weather forecast (as shown in Figure 19.1), to unstructured perceptual data,
10531 such as a raw image or video; the output may be a single sentence, such as an image cap-
10532 tion, or a multi-paragraph argument. Despite this diversity of conditions, all data-to-text
10533 systems share some of the same challenges (Reiter and Dale, 2000):

- 10534 • determining what parts of the data to describe;
10535 • planning a presentation of this information;
10536 • **lexicalizing** the data into words and phrases;
10537 • organizing words and phrases into well-formed sentences and paragraphs.

10538 The earlier stages of this process are sometimes called **content selection** and **text plan-**
10539 **ning**; the later stages are often called **surface realization**.

10540 Early systems for data-to-text generation were modular, with separate software com-
10541 ponents for each task. Artificial intelligence **planning** algorithms can be applied to both

Temperature				Cloud sky cover	
<i>time</i>	<i>min</i>	<i>mean</i>	<i>max</i>	<i>time</i>	<i>percent (%)</i>
06:00-21:00	9	15	21	06:00-09:00	25-50
Wind speed				Wind direction	
<i>time</i>	<i>min</i>	<i>mean</i>	<i>max</i>	<i>time</i>	<i>mode</i>
06:00-21:00	15	20	30	06:00-21:00	S

Cloudy, with temperatures between 10 and 20 degrees. South wind around 20 mph.

Figure 19.1: An example input-output pair for the task of generating text descriptions of weather forecasts (adapted from Konstas and Lapata, 2013).

10542 the high-level information structure and the organization of individual sentences, ensur-
 10543 ing that communicative goals are met (McKeown, 1992; Moore and Paris, 1993). Surface
 10544 realization can be performed by grammars or templates, which link specific types of data
 10545 to candidate words and phrases. A simple example template is offered by Wiseman et al.
 10546 (2017), for generating descriptions of basketball games:

- 10547 (19.1) The <team1>(<wins1>-<losses1>) defeated the <team2>(<wins2>-<losses2>),
 10548 <pts1>-<pts2>.
 10549 The New York Knicks (45-5) defeated the Boston Celtics (11-38), 115-79.

10550 For more complex cases, it may be necessary to apply morphological inflections such as
 10551 pluralization and tense marking — even in the simple example above, languages such
 10552 as Russian would require case marking suffixes for the team names. Such inflections can
 10553 be applied as a postprocessing step. Another difficult challenge for surface realization is
 10554 the generation of varied **referring expressions** (e.g., *The Knicks, New York, they*), which is
 10555 critical to avoid repetition. As discussed in § 16.2.1, the form of referring expressions is
 10556 constrained by the discourse and information structure.

10557 An example at the intersection of rule-based and statistical techniques is the NITRO-
 10558 GEN system (Langkilde and Knight, 1998). The input to NITROGEN is an abstract meaning
 10559 representation (AMR; see § 13.3) of semantic content to be expressed in a single sentence.
 10560 In data-to-text scenarios, the abstract meaning representation is the output of a higher-
 10561 level text planning stage. A set of rules then converts the abstract meaning representation
 10562 into various sentence plans, which may differ in both the high-level structure (e.g., active
 10563 versus passive voice) as well as the low-level details (e.g., word and phrase choice). Some
 10564 examples are shown in Figure 19.2. To control the combinatorial explosion in the number
 10565 of possible realizations for any given meaning, the sentence plans are unified into a single
 10566 finite-state acceptor, in which word tokens are represented by arcs (see § 9.1.1). A bigram

```
(a / admire-01
 :ARG0 (v / visitor
        :ARG1-of (c / arrive-01
                   :ARG4 (j / Japan)))
        :ARG1 (m / "Mount Fuji"))
```

- Visitors who came to Japan admire Mount Fuji.
- Visitors who came in Japan admire Mount Fuji.
- Mount Fuji is admired by the visitor who came in Japan.

Figure 19.2: Abstract meaning representation and candidate surface realizations from the NITROGEN system. Example adapted from Langkilde and Knight (1998).

language model is then used to compute weights on the arcs, so that the shortest path is also the surface realization with the highest bigram language model probability.

More recent systems are unified models that are trained end-to-end using backpropagation. Data-to-text generation shares many properties with machine translation, including a problem of **alignment**: labeled examples provide the data and the text, but they do not specify which parts of the text correspond to which parts of the data. For example, to learn from Figure 19.1, the system must align the word *cloudy* to records in CLOUD SKY COVER, the phrases *10* and *20 degrees* to the MIN and MAX fields in TEMPERATURE, and so on. As in machine translation, both latent variables and neural attention have been proposed as solutions.

19.1.1 Latent data-to-text alignment

Given a dataset of texts and associated records $\{(\mathbf{w}^{(i)}, \mathbf{y}^{(i)})\}_{i=1}^N$, our goal is to learn a model Ψ , so that

$$\hat{\mathbf{w}} = \underset{\mathbf{w} \in \mathcal{V}^*}{\operatorname{argmax}} \Psi(\mathbf{w}, \mathbf{y}; \theta), \quad [19.1]$$

where \mathcal{V}^* is the set of strings over a discrete vocabulary, and θ is a vector of parameters. The relationship between \mathbf{w} and \mathbf{y} is complex: the data \mathbf{y} may contain dozens of records, and \mathbf{w} may extend to several sentences. To facilitate learning and inference, it would be helpful to decompose the scoring function Ψ into subcomponents. This would be possible if given an **alignment**, specifying which element of \mathbf{y} is expressed in each part of \mathbf{w} . Specifically, let z_m indicates the record aligned to word m . For example, in Figure 19.1, z_1 might specify that the word *cloudy* is aligned to the record *cloud-sky-cover:percent*. The score for this alignment would then be given by the weight on features such as

$$(\textit{cloudy}, \textit{cloud-sky-cover:percent}). \quad [19.2]$$

In general, given an observed set of alignments, the score for a generation can be

10589 written as sum of local scores (Angeli et al., 2010):

$$\Psi(\mathbf{w}, \mathbf{y}; \theta) = \sum_{m=1}^M \psi_{w,y}(\mathbf{w}_m, \mathbf{y}_{z_m}) + \psi_w(w_m, w_{m-1}) + \psi_z(z_m, z_{m-1}), \quad [19.3]$$

10590 where ψ_w can represent a bigram language model, and ψ_z can be tuned to reward coherence,
 10591 such as the use of related records in nearby words.¹ The parameters of this model
 10592 could be learned from labeled data $\{(\mathbf{w}^{(i)}, \mathbf{y}^{(i)}, \mathbf{z}^{(i)})\}_{i=1}^N$. However, while several datasets
 10593 include structured records and natural language text (Barzilay and McKeown, 2005; Chen
 10594 and Mooney, 2008; Liang and Klein, 2009), the alignments between text and records are
 10595 usually not available.² One solution is to model the problem probabilistically, treating the
 10596 alignment as a latent variable (Liang et al., 2009; Konstas and Lapata, 2013). The model
 10597 can then be estimated using expectation maximization or sampling (see chapter 5).

10598 19.1.2 Neural data-to-text generation

10599 The **encoder-decoder model** and **neural attention** were introduced in § 18.3 as methods
 10600 for neural machine translation. They can also be applied to data-to-text generation, with
 10601 the data acting as the source language (Mei et al., 2016). In neural machine translation,
 10602 the attention mechanism linked words in the source to words in the target; in data-to-
 10603 text generation, the attention mechanism can link each part of the generated text back
 10604 to a record in the data. The biggest departure from translation is in the encoder, which
 10605 depends on the form of the data.

10606 Data encoders

10607 In some types of structured records, all values are drawn from discrete sets. For example,
 10608 the birthplace of an individual is drawn from a discrete set of possible locations; the diag-
 10609 nosis and treatment of a patient are drawn from an exhaustive list of clinical codes (John-
 10610 son et al., 2016). In such cases, vector embeddings can be estimated for each field and
 10611 possible value: for example, a vector embedding for the field BIRTHPLACE, and another
 10612 embedding for the value BERKELEY_CALIFORNIA (Bordes et al., 2011). The table of such
 10613 embeddings serves as the encoding of a structured record (He et al., 2017). It is also possi-
 10614 ble to compress the entire table into a single vector representation, by **pooling** across the
 10615 embeddings of each field and value (Lebret et al., 2016).

¹More expressive decompositions of Ψ are possible. For example, Wong and Mooney (2007) use a synchronous context-free grammar (see § 18.2.4) to “translate” between a meaning representation and natural language text.

²An exception is a dataset of records and summaries from American football games, containing annotations of alignments between sentences and records (Snyder and Barzilay, 2007).

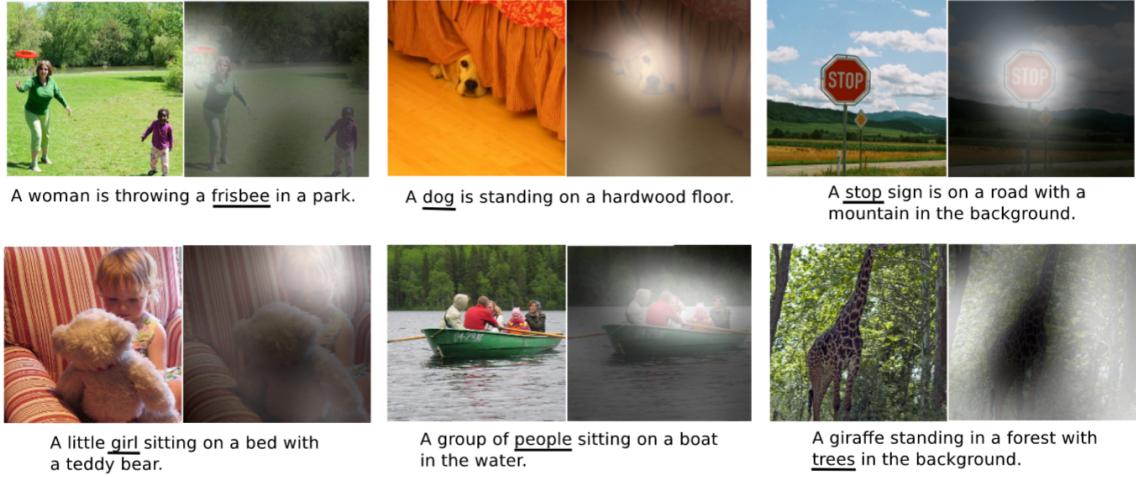


Figure 19.3: Examples of the image captioning task, with attention masks shown for each of the underlined words (Xu et al., 2015).

Sequences Some types of structured records have a natural ordering, such as events in a game (Chen and Mooney, 2008) and steps in a recipe (Tutin and Kittredge, 1992). For example, the following records describe a sequence of events in a robot soccer match (Mei et al., 2016):

```
PASS(arg1 = PURPLE6, arg2 = PURPLE3)
KICK(arg1 = PURPLE3)
BADPASS(arg1 = PURPLE3, arg2 = PINK9).
```

10616 Each event is a single record, and can be encoded by a concatenation of vector representations for the event type (e.g., PASS), the field (e.g., arg1), and the values (e.g., PURPLE3),
10617 e.g.,
10618

$$\mathbf{X} = [\mathbf{u}_{\text{PASS}}, \mathbf{u}_{\text{arg1}}, \mathbf{u}_{\text{PURPLE6}}, \mathbf{u}_{\text{arg2}}, \mathbf{u}_{\text{PURPLE3}}]. \quad [19.4]$$

10619 This encoding can then act as the input layer for a recurrent neural network, yielding a
10620 sequence of vector representations $\{\mathbf{z}_r\}_{r=1}^R$, where r indexes over records. Interestingly,
10621 this sequence-based approach can work even in cases where there is no natural ordering
10622 over the records, such as the weather data in Figure 19.1 (Mei et al., 2016).

10623 **Images** Another flavor of data-to-text generation is the generation of text captions for
10624 images. Examples from this task are shown in Figure 19.3. Images are naturally represented
10625 as tensors: a color image of 320×240 pixels would be stored as a tensor with
10626 $320 \times 240 \times 3$ intensity values. The dominant approach to image classification is to encode
10627 images as vectors using a combination of convolution and pooling (Krizhevsky et al.,

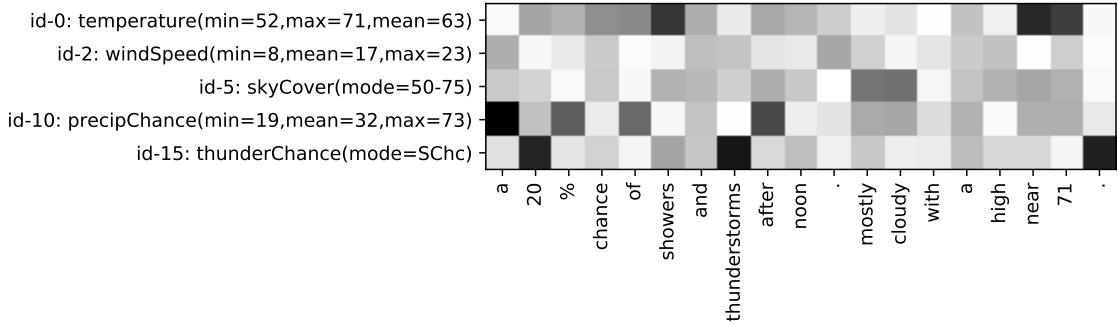


Figure 19.4: Neural attention in text generation. Figure adapted from Mei et al. (2016).

10628 2012). Chapter 3 explains how to use convolutional networks for text; for images, convolution
 10629 is applied across the vertical, horizontal, and color dimensions. By pooling the re-
 10630 sults of successive convolutions, the image is converted to a vector representation, which
 10631 can then be fed directly into the decoder as the initial state (Vinyals et al., 2015), just as
 10632 in the sequence-to-sequence translation model (see § 18.3). Alternatively, one can apply
 10633 a set of convolutional networks, yielding vector representations for different parts of the
 10634 image, which can then be combined using neural attention (Xu et al., 2015).

10635 Attention

Given a set of embeddings of the data $\{\mathbf{z}_r\}_{r=1}^R$ and a decoder state \mathbf{h}_m , an attention vector over the data can be computed using the same techniques as in machine translation (see § 18.3.1). When generating word m of the output, attention is computed over the records,

$$\psi_\alpha(m, r) = \beta_\alpha \cdot f(\Theta_\alpha[\mathbf{h}_m; \mathbf{z}_r]) \quad [19.5]$$

$$\boldsymbol{\alpha}_m = g([\psi_\alpha(m, 1), \psi_\alpha(m, 2), \dots, \psi_\alpha(m, R)]) \quad [19.6]$$

$$\mathbf{c}_m = \sum_{r=1}^R \alpha_{m \rightarrow r} \mathbf{z}_r, \quad [19.7]$$

10636 where f is an elementwise nonlinearity such as tanh or ReLU, and g is either softmax or
 10637 elementwise sigmoid. The weighted sum \mathbf{c}_m can then be included in the recurrent update
 10638 to the decoder state, or in the emission probabilities, as described in § 18.3.1. Figure 19.4
 10639 shows the attention to components of a weather record, while generating the text shown
 10640 on the x -axis.

10641 Adapting this architecture to image captioning is straightforward. A convolutional
 10642 neural networks is applied to a set of image locations, and the output at each location ℓ is
 10643 represented with a vector \mathbf{z}_ℓ . Attention can then be computed over the image locations,
 10644 as shown in the right panels of each pair of images in Figure 19.3.

Various modifications to this basic mechanism have been proposed. In **coarse-to-fine attention** (Mei et al., 2016), each record receives a global attention $a_r \in [0, 1]$, which is independent of the decoder state. This global attention, which represents the overall importance of the record, is multiplied with the decoder-based attention scores, before computing the final normalized attentions. In **structured attention**, the attention vector $\alpha_{m \rightarrow \cdot}$ can include structural biases, which can favor assigning higher attention values to contiguous segments or to dependency subtrees (Kim et al., 2017). Structured attention vectors can be computed by running the forward-backward algorithm to obtain marginal attention probabilities (see § 7.5.3). Because each step in the forward-backward algorithm is differentiable, it can be encoded in a computation graph, and end-to-end learning can be performed by backpropagation.

Decoder

Given the encoding, the decoder can function just as in neural machine translation (see § 18.3.1), using the attention-weighted encoder representation in the decoder recurrence and/or output computation. As in machine translation, beam search can help to avoid search errors (Lebret et al., 2016).

Many applications require generating words that do not appear in the training vocabulary. For example, a weather record may contain a previously unseen city name; a sports record may contain a previously unseen player name. Such tokens can be generated in the text by copying them over from the input (e.g., Gulcehre et al., 2016).³ First introduce an additional variable $s_m \in \{\text{gen}, \text{copy}\}$, indicating whether token $w_m^{(t)}$ should be generated or copied. The decoder probability is then,

$$p(w^{(t)} | w_{1:m-1}^{(t)}, \mathbf{Z}, s_m) = \begin{cases} \text{SoftMax}(\beta_{w^{(t)}} \cdot \mathbf{h}_{m-1}^{(t)}), & s_m = \text{gen} \\ \sum_{r=1}^R \delta(w_r^{(s)} = w^{(t)}) \times \alpha_{m \rightarrow r}, & s_m = \text{copy}, \end{cases} \quad [19.8]$$

where $\delta(w_r^{(s)} = w^{(t)})$ is an indicator function, taking the value 1 iff the text of the record $w_r^{(s)}$ is identical to the target word $w^{(t)}$. The probability of copying record r from the source is $\delta(s_m = \text{copy}) \times \alpha_{m \rightarrow r}$, the product of the copy probability by the local attention. Note that in this model, the attention weights α_m are computed from the *previous* decoder state \mathbf{h}_{m-1} . The computation graph therefore remains a feedforward network, with recurrent paths such as $\mathbf{h}_{m-1}^{(t)} \rightarrow \alpha_m \rightarrow w_m^{(t)} \rightarrow \mathbf{h}_m^{(t)}$.

To facilitate end-to-end training, the switching variable s_m can be represented by a gate π_m , which is computed from a two-layer feedforward network, whose input consists of the concatenation of the decoder state $\mathbf{h}_{m-1}^{(t)}$ and the attention-weighted representation

³A number of variants of this strategy have been proposed (e.g., Gu et al., 2016; Merity et al., 2017). See Wiseman et al. (2017) for an overview.

10670 of the data, $\mathbf{c}_m = \sum_{r=1}^R \alpha_{m \rightarrow r} \mathbf{z}_r$,

$$\pi_m = \sigma(\Theta^{(2)} f(\Theta^{(1)}[\mathbf{h}_{m-1}^{(t)}; \mathbf{c}_m])). \quad [19.9]$$

The full generative probability at token m is then,

$$\begin{aligned} p(w^{(t)} | \mathbf{w}_{1:m}^{(t)}, \mathbf{Z}) = & \pi_m \times \underbrace{\frac{\exp \beta_{w^{(t)}} \cdot \mathbf{h}_{m-1}^{(t)}}{\sum_{j=1}^V \exp \beta_j \cdot \mathbf{h}_{m-1}^{(t)}}}_{\text{generate}} + (1 - \pi_m) \times \underbrace{\sum_{r=1}^R \delta(w_r^{(s)} = w^{(t)}) \times \alpha_{m \rightarrow r}}_{\text{copy}}. \end{aligned} \quad [19.10]$$

10671 19.2 Text-to-text generation

10672 Text-to-text generation includes problems of summarization and simplification:

- 10673 • reading a novel and outputting a paragraph-long summary of the plot;⁴
- 10674 • reading a set of blog posts about politics, and outputting a bullet list of the various
- 10675 issues and perspectives;
- 10676 • reading a technical research article about the long-term health consequences of drink-
- 10677 ing kombucha, and outputting a summary of the article in language that non-experts
- 10678 can understand.

10679 These problems can be approached in two ways: through the encoder-decoder architec-
10680 ture discussed in the previous section, or by operating directly on the input text.

10681 19.2.1 Neural abstractive summarization

10682 **Sentence summarization** is the task of shortening a sentence while preserving its mean-
10683 ing, as in the following examples (Knight and Marcu, 2000; Rush et al., 2015):

- 10684 (19.2) a. The documentation is typical of Epson quality: excellent.
10685 Documentation is excellent.
- 10686 b. Russian defense minister Ivanov called sunday for the creation of a joint front
10687 for combating global terrorism.
10688 Russia calls for joint front against terrorism.

⁴In § 16.3.4, we encountered a special case of single-document summarization, which involved extracting the most important sentences or discourse units. We now consider the more challenging problem of **abstractive summarization**, in which the summary can include words that do not appear in the original text.

10691 Sentence summarization is closely related to **sentence compression**, in which the sum-
 10692 mary is produced by deleting words or phrases from the original (Clarke and Lapata,
 10693 2008). But as shown in (19.2b), a sentence summary can also introduce new words, such
 10694 as *against*, which replaces the phrase *for combatting*.

10695 Sentence summarization can be treated as a machine translation problem, using the at-
 10696 tentional encoder-decoder translation model discussed in § 18.3.1 (Rush et al., 2015). The
 10697 longer sentence is encoded into a sequence of vectors, one for each token. The decoder
 10698 then computes attention over these vectors when updating its own recurrent state. As
 10699 with data-to-text generation, it can be useful to augment the encoder-decoder model with
 10700 the ability to copy words directly from the source. Rush et al. (2015) train this model by
 10701 building four million sentence pairs from news articles. In each pair, the longer sentence is
 10702 the first sentence of the article, and the summary is the article headline. Sentence summa-
 10703 rization can also be trained in a semi-supervised fashion, using a probabilistic formulation
 10704 of the encoder-decoder model called a **variational autoencoder** (Miao and Blunsom, 2016,
 10705 also see § 14.8.2).

When summarizing longer documents, an additional concern is that the summary not
 be repetitive: each part of the summary should cover new ground. This can be addressed
 by maintaining a vector of the sum total of all attention values thus far, $t_m = \sum_{n=1}^m \alpha_n$.
 This total can be used as an additional input to the computation of the attention weights,

$$\alpha_{m \rightarrow n} \propto \exp \left(\mathbf{v}_\alpha \cdot \tanh(\Theta_\alpha[\mathbf{h}_m^{(t)}; \mathbf{h}_n^{(s)}; \mathbf{t}_m]) \right), \quad [19.11]$$

which enables the model to learn to prefer parts of the source which have not been at-
 tended to yet (Tu et al., 2016). To further encourage diversity in the generated summary,
 See et al. (2017) introduce a **coverage loss** to the objective function,

$$\ell_m = \sum_{n=1}^{M^{(s)}} \min(\alpha_{m \rightarrow n}, t_{m \rightarrow n}). \quad [19.12]$$

10706 This loss will be low if α_m assigns little attention to words that already have large values in
 10707 t_m . Coverage loss is similar to the concept of **marginal relevance**, in which the reward for
 10708 adding new content is proportional to the extent to which it increases the overall amount
 10709 of information conveyed by the summary (Carbonell and Goldstein, 1998).

10710 19.2.2 Sentence fusion for multi-document summarization

10711 In **multi-document summarization**, the goal is to produce a summary that covers the
 10712 content of several documents (McKeown et al., 2002). One approach to this challenging
 10713 problem is to identify sentences across multiple documents that relate to a single theme,
 10714 and then to fuse them into a single sentence (Barzilay and McKeown, 2005). As an exam-
 10715 ple, consider the following two sentences (McKeown et al., 2010):

- 10716 (19.3) a. Palin actually turned against the bridge project only after it became a national
 10717 symbol of wasteful spending.
 10718 b. Ms. Palin supported the bridge project while running for governor, and
 10719 abandoned it after it became a national scandal.

10720 An *intersection* preserves only the content that is present in both sentences:

- 10721 (19.4) Palin turned against the bridge project after it became a national scandal.

10722 A *union* includes information from both sentences:

- 10723 (19.5) Ms. Palin supported the bridge project while running for governor, but turned
 10724 against it when it became a national scandal and a symbol of wasteful spending.

Dependency parsing is often used as a technique for sentence fusion. After parsing each sentence, the resulting dependency trees can be aggregated into a lattice (Barzilay and McKeown, 2005) or a graph structure (Filippova and Strube, 2008), in which identical or closely related words (e.g., *Palin*, *bridge*, *national*) are fused into a single node. The resulting graph can then be pruned back to a tree by solving an **integer linear program** (see § 13.2.2),

$$\max_{\mathbf{y}} \sum_{i,j,r} \psi(i \xrightarrow{r} j, \mathbf{w}; \boldsymbol{\theta}) \times y_{i,j,r} \quad [19.13]$$

$$\text{s.t. } \mathbf{y} \in \mathcal{C}, \quad [19.14]$$

10725 where the variable $y_{i,j,r} \in \{0, 1\}$ indicates whether there is an edge from i to j of type r ,
 10726 the score of this edge is $\psi(i \xrightarrow{r} j, \mathbf{w}; \boldsymbol{\theta})$, and \mathcal{C} is a set of constraints, which ensures that \mathbf{y}
 10727 forms a valid dependency graph. As usual, \mathbf{w} is the list of words in the graph, and $\boldsymbol{\theta}$ is a
 10728 vector of parameters. The score $\psi(i \xrightarrow{r} j, \mathbf{w}; \boldsymbol{\theta})$ reflects the “importance” of the modifier
 10729 j to the overall meaning: in intersective fusion, this score indicates the extent to which
 10730 the content in this edge is expressed in all sentences; in union fusion, the score indicates
 10731 whether the content in the edge is expressed in any sentence. The constraint set \mathcal{C} can
 10732 impose additional linguistic constraints: for example, ensuring that coordinated nouns
 10733 are sufficiently similar. The resulting tree must then be **linearized** into a sentence. Lin-
 10734 earization is like the inverse of dependency parsing: instead of parsing from a sequence
 10735 of tokens into a tree, we must convert the tree back into a sequence of tokens. This is
 10736 typically done by generating a set of candidate linearizations, and choosing the one with
 10737 the highest score under a language model (Langkilde and Knight, 1998; Song et al., 2016).

10738 19.3 Dialogue

10739 **Dialogue systems** are capable of conversing with a human interlocutor, often to per-
 10740 form some task (Grosz, 1979), but sometimes just to chat (Weizenbaum, 1966). While re-

- (19.6) A: I want to order a pizza.
 B: What toppings?
 A: Anchovies.
 B: Ok, what address?
 A: The College of Computing building.
 B: Please confirm: one pizza with artichokes, to be delivered to the College of Computing building.
 A: No.
 B: What toppings?
 ...

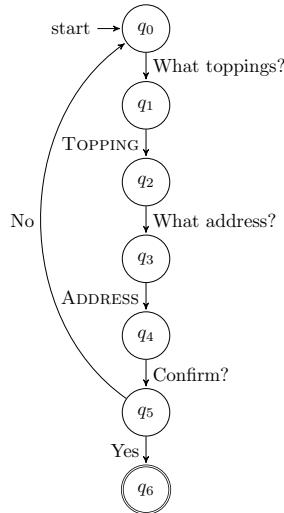


Figure 19.5: An example dialogue and the associated finite-state model. In the finite-state model, SMALL CAPS indicates that the user must provide information of this type in their answer.

10741 search on dialogue systems goes back several decades (Carbonell, 1970; Winograd, 1972),
 10742 commercial systems such as Alexa and Siri have recently brought this technology into
 10743 widespread use. Nonetheless, there is a significant gap between research and practice:
 10744 many practical dialogue systems remain scripted and inflexible, while research systems
 10745 emphasize abstractive text generation, “on-the-fly” decision making, and probabilistic
 10746 reasoning about the user’s intentions.

10747 19.3.1 Finite-state and agenda-based dialogue systems

10748 Finite-state automata were introduced in chapter 9 as a formal model of computation,
 10749 in which string inputs and outputs are linked to transitions between a finite number of
 10750 discrete states. This model naturally fits simple task-oriented dialogues, such as the one
 10751 shown in the left panel of Figure 19.5. This (somewhat frustrating) dialogue can be repre-
 10752 sented with a finite-state transducer, as shown in the right panel of the figure. The accept-
 10753 ing state is reached only when the two needed pieces of information are provided, and the
 10754 human user confirms that the order is correct. In this simple scenario, the TOPPING and
 10755 ADDRESS are the two **slots** associated with the activity of ordering a pizza, which is called
 10756 a **frame**. Frame representations can be hierarchical: for example, an ADDRESS could have
 10757 slots of its own, such as STREET and CITY.

10758 In the example dialogue in Figure 19.5, the user provides the precise inputs that are
 10759 needed in each turn (e.g., *anchovies*; *the College of Computing building*). Some users may

10760 prefer to communicate more naturally, with phrases like *I'd, uh, like some anchovies please.*
 10761 One approach to handling such utterances is to design a custom grammar, with non-
 10762 terminals for slots such as TOPPING and LOCATION. However, context-free parsing of
 10763 unconstrained speech input is challenging. A more lightweight alternative is BIO-style
 10764 sequence labeling (see § 8.3), e.g.:

10765 (19.7) *I'd like anchovies , and please bring it to the College of Computing*
 O O B-TOPPING O O O O O O B-ADDR I-ADDR I-ADDR I-ADDR
 10766 *Building .*
 I-ADDR O

10767 The tagger can be driven by a bi-directional recurrent neural network, similar to recurrent
 10768 approaches to semantic role labeling described in § 13.2.3.

10769 The input in (19.7) could not be handled by the finite-state system from Figure 19.5,
 10770 which forces the user to provide the topping first, and then the location. In this sense, the
 10771 “initiative” is driven completely by the system. **Agenda-based dialogue systems** extend
 10772 finite-state architectures by attempting to recognize all slots that are filled by the user’s re-
 10773 ply, thereby handling these more complex examples. Agenda-based systems dynamically
 10774 pose additional questions until the frame is complete (Bobrow et al., 1977; Allen et al.,
 10775 1995; Rudnicky and Xu, 1999). Such systems are said to be **mixed-initiative**, because both
 10776 the user and the system can drive the direction of the dialogue.

10777 19.3.2 Markov decision processes

10778 The task of dynamically selecting the next move in a conversation is known as **dialogue**
 10779 **management**. This problem can be framed as a **Markov decision process**, which is a
 10780 theoretical model that includes a discrete set of states, a discrete set of actions, a function
 10781 that computes the probability of transitions between states, and a function that computes
 10782 the cost or reward of action-state pairs. Let’s see how each of these elements pertains to
 10783 the pizza ordering dialogue system.

- 10784 • Each state is a tuple of information about whether the topping and address are
 10785 known, and whether the order has been confirmed. For example,

(KNOWN TOPPING, UNKNOWN ADDRESS, NOT CONFIRMED) [19.15]

10786 is a possible state. Any state in which the pizza order is confirmed is a terminal
 10787 state, and the Markov decision process stops after entering such a state.

- 10788 • The set of actions includes querying for the topping, querying for the address, and
 10789 requesting confirmation. Each action induces a probability distribution over states,
 10790 $p(s_t | a_t, s_{t-1})$. For example, requesting confirmation of the order is not likely to

10791 result in a transition to the terminal state if the topping is not yet known. This
 10792 probability distribution over state transitions may be learned from data, or it may
 10793 be specified in advance.

- 10794 • Each state-action-state tuple earns a reward, $r_a(s_t, s_{t+1})$. In the context of the pizza
 10795 ordering system, a simple reward function would be,

$$r_a(s_t, s_{t-1}) = \begin{cases} 0, & a = \text{CONFIRM}, s_t = (*, *, \text{CONFIRMED}) \\ -10, & a = \text{CONFIRM}, s_t = (*, *, \text{NOT CONFIRMED}) \\ -1, & a \neq \text{CONFIRM} \end{cases} \quad [19.16]$$

10796 This function assigns zero reward for successful transitions to the terminal state, a
 10797 large negative reward to a rejected request for confirmation, and a small negative re-
 10798 ward for every other type of action. The system is therefore rewarded for reaching
 10799 the terminal state in few steps, and penalized for prematurely requesting confirma-
 10800 tion.

10801 In a Markov decision process, a **policy** is a function $\pi : \mathcal{S} \rightarrow \mathcal{A}$ that maps from states
 10802 to actions (see § 15.2.4). The value of a policy is the expected sum of discounted rewards,
 10803 $E_\pi[\sum_{t=1}^T \gamma^t r_{a_t}(s_t, s_{t+1})]$, where γ is the discount factor, $\gamma \in [0, 1)$. Discounting has the
 10804 effect of emphasizing rewards that can be obtained immediately over less certain rewards
 10805 in the distant future.

10806 An optimal policy can be obtained by dynamic programming, by iteratively updating
 10807 the **value function** $V(s)$, which is the expectation of the cumulative reward from s under
 10808 the optimal action a ,

$$V(s) \leftarrow \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} p(s' | s, a)[r_a(s, s') + \gamma V(s')]. \quad [19.17]$$

10809 The value function $V(s)$ is computed in terms of $V(s')$ for all states $s' \in \mathcal{S}$. A series
 10810 of iterative updates to the value function will eventually converge to a stationary point.
 10811 This algorithm is known as **value iteration**. Given the converged value function $V(s)$, the
 10812 optimal action at each state is the argmax,

$$\pi(s) = \operatorname{argmax}_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} p(s' | s, a)[r_a(s, s') + \gamma V(s')]. \quad [19.18]$$

10813 Value iteration and related algorithms are described in detail by Sutton and Barto (1998).
 10814 For applications to dialogue systems, see Levin et al. (1998) and Walker (2000).

10815 The Markov decision process framework assumes that the current state of the dialogue
 10816 is known. In reality, the system may misinterpret the user's statements — for example,
 10817 believing that a specification of the delivery location (PEACHTREE) is in fact a specification

of the topping (PEACHES). In a **partially observable Markov decision process (POMDP)**, the system receives an *observation* o , which is probabilistically conditioned on the state, $p(o | s)$. It must therefore maintain a distribution of beliefs about which state it is in, with $q_t(s)$ indicating the degree of belief that the dialogue is in state s at time t . The POMDP formulation can help to make dialogue systems more robust to errors, particularly in the context of spoken language dialogues, where the speech itself may be misrecognized (Roy et al., 2000; Williams and Young, 2007). However, finding the optimal policy in a POMDP is computationally intractable, requiring additional approximations.

19.3.3 Neural chatbots

It's easier to talk when you don't need to get anything done. **Chatbots** are systems that parry the user's input with a response that keeps the conversation going. They can be built from the encoder-decoder architecture discussed in § 18.3 and § 19.1.2: the encoder converts the user's input into a vector, and the decoder produces a sequence of words as a response. For example, Shang et al. (2015) apply the attentional encoder-decoder translation model, training on a dataset of posts and responses from the Chinese microblogging platform Sina Weibo.⁵ This approach is capable of generating replies that relate thematically to the input, as shown in the following examples (translated from Chinese by Shang et al., 2015).

- 10836 (19.8) a. A: High fever attacks me every New Year's day.
10837 B: Get well soon and stay healthy!
- 10838 b. A: I gain one more year. Grateful to my group, so happy.
10839 B: Getting old now. Time has no mercy.

While encoder-decoder models can generate responses that make sense in the context of the immediately preceding turn, they struggle to maintain coherence over longer conversations. One solution is to model the dialogue context recurrently. This creates a **hierarchical recurrent network**, including both word-level and turn-level recurrences. The turn-level hidden state is then used as additional context in the decoder (Serban et al., 2016).

An open question is how to integrate the encoder-decoder architecture into task-oriented dialogue systems. Neural chatbots can be trained end-to-end: the user's turn is analyzed by the encoder, and the system output is generated by the decoder. This architecture can be trained by log-likelihood using backpropagation (e.g., Sordoni et al., 2015; Serban et al., 2016), or by more elaborate objectives, using reinforcement learning (Li et al., 2016). In contrast, the task-oriented dialogue systems described in § 19.3.1 typically involve a

⁵Twitter is also frequently used for construction of dialogue datasets (Ritter et al., 2011; Sordoni et al., 2015). Another source is technical support chat logs from the Ubuntu linux distribution (Uthus and Aha, 2013; Lowe et al., 2015).

10852 set of specialized modules: one for recognizing the user input, another for deciding what
10853 action to take, and a third for arranging the text of the system output.

10854 Recurrent neural network decoders can be integrated into Markov Decision Process
10855 dialogue systems, by conditioning the decoder on a representation of the information
10856 that is to be expressed in each turn (Wen et al., 2015). Specifically, the long short-term
10857 memory (LSTM; § 6.3) architecture is augmented so that the memory cell at turn m takes
10858 an additional input d_m , which is a representation of the slots and values to be expressed
10859 in the next turn. However, this approach still relies on additional modules to recognize
10860 the user’s utterance and to plan the overall arc of the dialogue.

10861 Another promising direction is to create embeddings for the elements in the domain:
10862 for example, the slots in a record and the entities that can fill them. The encoder then
10863 encodes not only the words of the user’s input, but the embeddings of the elements that
10864 the user mentions. Similarly, the decoder is endowed with the ability to refer to specific
10865 elements in the knowledge base. He et al. (2017) show that such a method can learn to
10866 play a collaborative dialogue game, in which both players are given a list of entities and
10867 their properties, and the goal is to find an entity that is on both players’ lists.

10868 Additional resources

10869 Gatt and Krahmer (2018) provide a comprehensive recent survey on text generation. For
10870 a book-length treatment of earlier work, see Reiter and Dale (2000). For a survey on image
10871 captioning, see Bernardi et al. (2016); for a survey of pre-neural approaches to dialogue
10872 systems, see Rieser and Lemon (2011). **Dialogue acts** were introduced in § 8.6 as a label-
10873 ing scheme for human-human dialogues; they also play a critical role in task-based dialogue
10874 systems (e.g., Allen et al., 1996). The incorporation of theoretical models of dialogue into
10875 computational systems is reviewed by Jurafsky and Martin (2009, chapter 24).

10876 While this chapter has focused on the informative dimension of text generation, an-
10877 other line of research aims to generate text with configurable stylistic properties (Walker
10878 et al., 1997; Mairesse and Walker, 2011; Ficler and Goldberg, 2017; Hu et al., 2017). This
10879 chapter also does not address the generation of creative text such as narratives (Riedl and
10880 Young, 2010), jokes (Ritchie, 2001), poems (Colton et al., 2012), and song lyrics (Gonçalo Oliveira
10881 et al., 2007).

10882 Exercises

- 10883 1. Find an article about a professional basketball game, with an associated “box score”
10884 of statistics. Which are the first three elements in the box score that are expressed
10885 in the article? Can you identify template-based patterns that express these elements
10886 of the record? Now find a second article about a different basketball game. Does it

mention the same first three elements of the box score? Do your templates capture how these elements are expressed in the text?

- 10889 2. This exercise is to be done by a pair of students. One student should choose an article
10890 from the news or from Wikipedia, and manually perform semantic role labeling
10891 (SRL) on three short sentences or clauses. (See chapter 13 for a review of SRL.)
10892 Identify the main the semantic relation and its arguments and adjuncts. Pass this
10893 structured record — but not the original sentence — to the other student, whose
10894 job is to generate a sentence expressing the semantics. Then reverse roles, and try
10895 to regenerate three sentences from another article, based on the predicate-argument
10896 semantics.
 - 10897 3. Compute the BLEU scores (see § 18.1.1) for the generated sentences in the previous
10898 problem, using the original article text as the reference.
 - 10899 4. Align each token in the text of Figure 19.1 to a specific single record in the database,
10900 or to the null record \emptyset . For example, the tokens *south wind* would align to the record
10901 *wind direction: 06:00-21:00: mode=S*. How often is each token aligned
10902 to the same record as the previous token? How many transitions are there? How
10903 might a system learn to output *10 degrees* for the record *min=9*?
 - 10904 5. In sentence compression and fusion, we may wish to preserve contiguous sequences
10905 of tokens (*n*-grams) and/or dependency edges. Find five short news articles with
10906 headlines. For each headline, compute the fraction of bigrams that appear in the
10907 main text of the article. Then do a manual dependency parse of the headline. For
10908 each dependency edge, count how often it appears as a dependency edge in the
10909 main text. You may use an automatic dependency parser to assist with this exercise,
10910 but check the output, and focus on UD 2.0 dependency grammar, as described in
10911 chapter 11.
 - 10912 6. § 19.2.2 presents the idea of generating text from dependency trees, which requires
10913 **linearization**. Sometimes there are multiple ways that a dependency tree can be
10914 linearized. For example:
- 10915 (19.9) a. The sick kids stayed at home in bed.
10916 b. The sick kids stayed in bed at home.

10917 Both sentences have an identical dependency parse: both *home* and *bed* are (oblique)
10918 dependents of *stayed*.

10919 Identify two more English dependency trees that can each be linearized in more than
10920 one way, and try to use a different pattern of variation in each tree. As usual, specify
10921 your trees in the Universal Dependencies 2 style, which is described in chapter 11.

7. In § 19.3.2, we considered a pizza delivery service. Let's simplify the problem to take-out, where it is only necessary to determine the topping and confirm the order. The state is a tuple in which the first element is T if the topping is specified and $?$ otherwise, and the second element is either YES or NO, depending on whether the order has been confirmed. The actions are TOPPING? (request information about the topping) and CONFIRM? (request confirmation). The state transition function is:

$$p(s_t | s_{t-1} = (?, \text{NO}), a = \text{TOPPING?}) = \begin{cases} 0.9, & s_t = (\text{T}, \text{NO}) \\ 0.1, & s_t = (?, \text{NO}). \end{cases} \quad [19.19]$$

$$p(s_t | s_{t-1} = (?, \text{NO}), a = \text{CONFIRM?}) = \begin{cases} 1, & s_t = (?, \text{NO}). \end{cases} \quad [19.20]$$

$$p(s_t | s_{t-1} = (\text{T}, \text{NO}), a = \text{TOPPING?}) = \begin{cases} 1, & s_t = (\text{T}, \text{NO}). \end{cases} \quad [19.21]$$

$$p(s_t | s_{t-1} = (\text{T}, \text{NO}), a = \text{CONFIRM?}) = \begin{cases} 0.9, & s_t = (\text{T}, \text{YES}) \\ 0.1, & s_t = (\text{T}, \text{NO}). \end{cases} \quad [19.22]$$

10922 Using the reward function defined in Equation 19.16, the discount $\gamma = 0.9$, and the
 10923 initialization $V(s) = 0$, execute three iterations of Equation 19.17. After these three
 10924 iterations, compute the optimal action in each state. You can assume that for the
 10925 terminal states, $V(*, \text{YES}) = 0$, so you only need to compute the values for non-
 10926 terminal states, $V(?, \text{NO})$ and $V(\text{T}, \text{NO})$.

- 10927 8. There are several toolkits that allow you to train encoder-decoder translation models
 10928 “out of the box”, such as FAIRSEQ (Gehring et al., 2017), XNMT (Neubig et al., 2018),
 10929 TENSOR2TENSOR (Vaswani et al., 2018), and OPENNMT (Klein et al., 2017).⁶ Use one
 10930 of these toolkits to train a chatbot dialogue system, using either the NPS dialogue
 10931 corpus that comes with NLTK (Forsyth and Martell, 2007), or, if you are feeling more
 10932 ambitious, the Ubuntu dialogue corpus (Lowe et al., 2015).

⁶<https://github.com/facebookresearch/fairseq>; <https://github.com/neulab/xnmt>;
<https://github.com/tensorflow/tensor2tensor>; <http://opennmt.net/>

10933 **Appendix A**

10934 **Probability**

10935 Probability theory provides a way to reason about random events. The sorts of random
10936 events that are typically used to explain probability theory include coin flips, card draws,
10937 and the weather. It may seem odd to think about the choice of a word as akin to the flip of
10938 a coin, particularly if you are the type of person to choose words carefully. But random or
10939 not, language has proven to be extremely difficult to model deterministically. Probability
10940 offers a powerful tool for modeling and manipulating linguistic data.

10941 Probability can be thought of in terms of **random outcomes**: for example, a single coin
10942 flip has two possible outcomes, heads or tails. The set of possible outcomes is the **sample**
10943 **space**, and a subset of the **sample space** is an **event**. For a sequence of two coin flips,
10944 there are four possible outcomes, $\{HH, HT, TH, TT\}$, representing the ordered sequences
10945 heads-head, heads-tails, tails-heads, and tails-tails. The event of getting exactly one head
10946 includes two outcomes: $\{HT, TH\}$.

10947 Formally, a probability is a function from events to the interval between zero and one:
10948 $\Pr : \mathcal{F} \rightarrow [0, 1]$, where \mathcal{F} is the set of possible events. An event that is certain has proba-
10949 bility one; an event that is impossible has probability zero. For example, the probability
10950 of getting fewer than three heads on two coin flips is one. Each outcome is also an event
10951 (a set with exactly one element), and for two flips of a fair coin, the probability of each
10952 outcome is,

$$\Pr(\{HH\}) = \Pr(\{HT\}) = \Pr(\{TH\}) = \Pr(\{TT\}) = \frac{1}{4}. \quad [\text{A.1}]$$

10953 **A.1 Probabilities of event combinations**

10954 Because events are sets of outcomes, we can use set-theoretic operations such as comple-
10955 ment, intersection, and union to reason about the probabilities of events and their combi-
10956 nations.

10957 For any event A , there is a **complement** $\neg A$, such that:

- 10958 • The probability of the union $A \cup \neg A$ is $\Pr(A \cup \neg A) = 1$;
 10959 • The intersection $A \cap \neg A = \emptyset$ is the empty set, and $\Pr(A \cap \neg A) = 0$.

10960 In the coin flip example, the event of obtaining a single head on two flips corresponds to
 10961 the set of outcomes $\{HT, TH\}$; the complement event includes the other two outcomes,
 10962 $\{TT, HH\}$.

10963 A.1.1 Probabilities of disjoint events

10964 When two events have an empty intersection, $A \cap B = \emptyset$, they are **disjoint**. The probabili-
 10965 ty of the union of two disjoint events is equal to the sum of their probabilities,

$$A \cap B = \emptyset \Rightarrow \Pr(A \cup B) = \Pr(A) + \Pr(B). \quad [A.2]$$

10966 This is the **third axiom of probability**, and it can be generalized to any countable sequence
 10967 of disjoint events.

In the coin flip example, this axiom can derive the probability of the event of getting a single head on two flips. This event is the set of outcomes $\{HT, TH\}$, which is the union of two simpler events, $\{HT, TH\} = \{HT\} \cup \{TH\}$. The events $\{HT\}$ and $\{TH\}$ are disjoint. Therefore,

$$\Pr(\{HT, TH\}) = \Pr(\{HT\} \cup \{TH\}) = \Pr(\{HT\}) + \Pr(\{TH\}) \quad [A.3]$$

$$= \frac{1}{4} + \frac{1}{4} = \frac{1}{2}. \quad [A.4]$$

10968 In the general, the probability of the union of two events is,

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B). \quad [A.5]$$

This can be seen visually in Figure A.1, and it can be derived from the third axiom of probability. Consider an event that includes all outcomes in B that are not in A , denoted as $B - (A \cap B)$. By construction, this event is disjoint from A . We can therefore apply the additive rule,

$$\Pr(A \cup B) = \Pr(A) + \Pr(B - (A \cap B)). \quad [A.6]$$

Furthermore, the event B is the union of two disjoint events: $A \cap B$ and $B - (A \cap B)$.

$$\Pr(B) = \Pr(B - (A \cap B)) + \Pr(A \cap B). \quad [A.7]$$

Reorganizing and substituting into Equation A.6 gives the desired result:

$$\Pr(B - (A \cap B)) = \Pr(B) - \Pr(A \cap B) \quad [A.8]$$

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B). \quad [A.9]$$

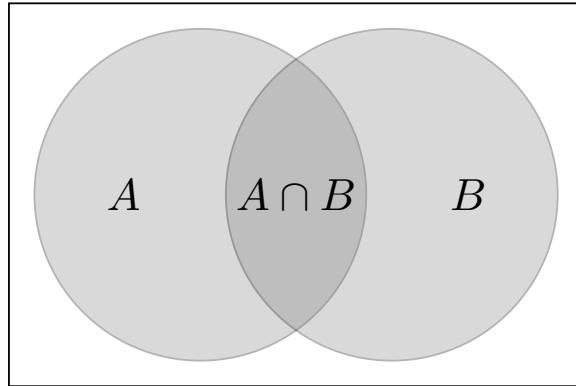


Figure A.1: A visualization of the probability of non-disjoint events A and B .

10969 **A.1.2 Law of total probability**

10970 A set of events $\mathcal{B} = \{B_1, B_2, \dots, B_N\}$ is a **partition** of the sample space iff each pair of
10971 events is disjoint ($B_i \cap B_j = \emptyset$), and the union of the events is the entire sample space.
10972 The law of total probability states that we can **marginalize** over these events as follows,

$$\Pr(A) = \sum_{B_n \in \mathcal{B}} \Pr(A \cap B_n). \quad [\text{A.10}]$$

10973 For any event B , the union $B \cup \neg B$ is a partition of the sample space. Therefore, a special
10974 case of the law of total probability is,

$$\Pr(A) = \Pr(A \cap B) + \Pr(A \cap \neg B). \quad [\text{A.11}]$$

10975 **A.2 Conditional probability and Bayes' rule**

A **conditional probability** is an expression like $\Pr(A \mid B)$, which is the probability of the event A , assuming that event B happens too. For example, we may be interested in the probability of a randomly selected person answering the phone by saying *hello*, conditioned on that person being a speaker of English. Conditional probability is defined as the ratio,

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)}. \quad [\text{A.12}]$$

The **chain rule of probability** states that $\Pr(A \cap B) = \Pr(A \mid B) \times \Pr(B)$, which is just

a rearrangement of terms from Equation A.12. The chain rule can be applied repeatedly:

$$\begin{aligned}\Pr(A \cap B \cap C) &= \Pr(A | B \cap C) \times \Pr(B \cap C) \\ &= \Pr(A | B \cap C) \times \Pr(B | C) \times \Pr(C).\end{aligned}$$

Bayes' rule (sometimes called Bayes' law or Bayes' theorem) gives us a way to convert between $\Pr(A | B)$ and $\Pr(B | A)$. It follows from the definition of conditional probability and the chain rule:

$$\Pr(A | B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{\Pr(B | A) \times \Pr(A)}{\Pr(B)} \quad [\text{A.13}]$$

10976 Each term in Bayes rule has a name, which we will occasionally use:

- 10977 • Pr(A) is the **prior**, since it is the probability of event A without knowledge about
10978 whether B happens or not.
- 10979 • Pr($B | A$) is the **likelihood**, the probability of event B given that event A has oc-
10980 curred.
- 10981 • Pr($A | B$) is the **posterior**, the probability of event A with knowledge that B has
10982 occurred.

10983 **Example** The classic examples for Bayes' rule involve tests for rare diseases, but Man-
10984 ning and Schütze (1999) reframe this example in a linguistic setting. Suppose that you are
10985 interested in a rare syntactic construction, such as *parasitic gaps*, which occur on average
10986 once in 100,000 sentences. Here is an example of a parasitic gap:

10987 (A.1) *Which class did you attend ... without registering for ...?*

10988 Lana Linguist has developed a complicated pattern matcher that attempts to identify
10989 sentences with parasitic gaps. It's pretty good, but it's not perfect:

- 10990 • If a sentence has a parasitic gap, the pattern matcher will find it with probability
10991 0.95. (This is the **recall**, which is one minus the **false negative rate**.)
- 10992 • If the sentence doesn't have a parasitic gap, the pattern matcher will wrongly say it
10993 does with probability 0.005. (This is the **false positive rate**, which is one minus the
10994 **precision**.)

10995 Suppose that Lana's pattern matcher says that a sentence contains a parasitic gap. What
10996 is the probability that this is true?

Let G be the event of a sentence having a parasitic gap, and T be the event of the test being positive. We are interested in the probability of a sentence having a parasitic gap given that the test is positive. This is the conditional probability $\Pr(G | T)$, and it can be computed by Bayes' rule:

$$\Pr(G | T) = \frac{\Pr(T | G) \times \Pr(G)}{\Pr(T)}. \quad [\text{A.14}]$$

10997 We already know both terms in the numerator: $\Pr(T | G)$ is the recall, which is 0.95; $\Pr(G)$
10998 is the prior, which is 10^{-5} .

10999 We are not given the denominator, but it can be computed using tools developed earlier
11000 in this section. First apply the law of total probability, using the partition $\{G, \neg G\}$:

$$\Pr(T) = \Pr(T \cap G) + \Pr(T \cap \neg G). \quad [\text{A.15}]$$

This says that the probability of the test being positive is the sum of the probability of a **true positive** ($T \cap G$) and the probability of a **false positive** ($T \cap \neg G$). The probability of each of these events can be computed using the chain rule:

$$\Pr(T \cap G) = \Pr(T | G) \times \Pr(G) = 0.95 \times 10^{-5} \quad [\text{A.16}]$$

$$\Pr(T \cap \neg G) = \Pr(T | \neg G) \times \Pr(\neg G) = 0.005 \times (1 - 10^{-5}) \approx 0.005 \quad [\text{A.17}]$$

$$\Pr(T) = \Pr(T \cap G) + \Pr(T \cap \neg G) \quad [\text{A.18}]$$

$$= 0.95 \times 10^{-5} + 0.005. \quad [\text{A.19}]$$

Plugging these terms into Bayes' rule gives the desired posterior probability,

$$\Pr(G | T) = \frac{\Pr(T | G) \Pr(G)}{\Pr(T)} \quad [\text{A.20}]$$

$$= \frac{0.95 \times 10^{-5}}{0.95 \times 10^{-5} + 0.005 \times (1 - 10^{-5})} \quad [\text{A.21}]$$

$$\approx 0.002. \quad [\text{A.22}]$$

11001 Lana's pattern matcher seems accurate, with false positive and false negative rates
11002 below 5%. Yet the extreme rarity of the phenomenon means that a positive result from the
11003 detector is most likely to be wrong.

11004 A.3 Independence

Two events are independent if the probability of their intersection is equal to the product of their probabilities: $\Pr(A \cap B) = \Pr(A) \times \Pr(B)$. For example, for two flips of a fair

coin, the probability of getting heads on the first flip is independent of the probability of getting heads on the second flip:

$$\Pr(\{HT, HH\}) = \Pr(HT) + \Pr(HH) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2} \quad [A.23]$$

$$\Pr(\{HH, TH\}) = \Pr(HH) + \Pr(TH) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2} \quad [A.24]$$

$$\Pr(\{HT, HH\}) \times \Pr(\{HH, TH\}) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} \quad [A.25]$$

$$\Pr(\{HT, HH\} \cap \{HH, TH\}) = \Pr(HH) = \frac{1}{4} \quad [A.26]$$

$$= \Pr(\{HT, HH\}) \times \Pr(\{HH, TH\}). \quad [A.27]$$

If $\Pr(A \cap B \mid C) = \Pr(A \mid C) \times \Pr(B \mid C)$, then the events A and B are **conditionally independent**, written $A \perp B \mid C$. Conditional independence plays a important role in probabilistic models such as Naïve Bayes chapter 2.

A.4 Random variables

Random variables are functions from events to \mathbb{R}^n , where \mathbb{R} is the set of real numbers. This subsumes several useful special cases:

- An **indicator random variable** is a function from events to the set $\{0, 1\}$. In the coin flip example, we can define Y as an indicator random variable, taking the value 1 when the coin has come up heads on at least one flip. This would include the outcomes $\{HH, HT, TH\}$. The probability $\Pr(Y = 1)$ is the sum of the probabilities of these outcomes, $\Pr(Y = 1) = \frac{1}{4} + \frac{1}{4} + \frac{1}{4} = \frac{3}{4}$.
- A **discrete random variable** is a function from events to a discrete subset of \mathbb{R} . Consider the coin flip example: the number of heads on two flips, X , can be viewed as a discrete random variable, $X \in \{0, 1, 2\}$. The event probability $\Pr(X = 1)$ can again be computed as the sum of the probabilities of the events in which there is one head, $\{HT, TH\}$, giving $\Pr(X = 1) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$.

Each possible value of a random variable is associated with a subset of the sample space. In the coin flip example, $X = 0$ is associated with the event $\{TT\}$, $X = 1$ is associated with the event $\{HT, TH\}$, and $X = 2$ is associated with the event $\{HH\}$. Assuming a fair coin, the probabilities of these events are, respectively, $1/4$, $1/2$, and $1/4$. This list of numbers represents the **probability distribution** over X , written p_X , which maps from the possible values of X to the non-negative reals. For a specific value x , we write $p_X(x)$, which is equal to the event probability $\Pr(X = x)$.¹ The function p_X is called

¹In general, capital letters (e.g., X) refer to random variables, and lower-case letters (e.g., x) refer to specific values. When the distribution is clear from context, I will simply write $p(x)$.

a probability **mass** function (pmf) if X is discrete; it is called a probability **density** function (pdf) if X is continuous. In either case, the function must sum to one, and all values must be non-negative:

$$\int_x p_X(x)dx = 1 \quad [A.28]$$

$$\forall x, p_X(x) \geq 0. \quad [A.29]$$

Probabilities over multiple random variables can written as **joint probabilities**, e.g., $p_{A,B}(a,b) = \Pr(A = a \cap B = b)$. Several properties of event probabilities carry over to probability distributions over random variables:

- The **marginal probability distribution** is $p_A(a) = \sum_b p_{A,B}(a,b)$.
- The **conditional probability distribution** is $p_{A|B}(a | b) = \frac{p_{A,B}(a,b)}{p_B(b)}$.
- Random variables A and B are independent iff $p_{A,B}(a,b) = p_A(a) \times p_B(b)$.

A.5 Expectations

Sometimes we want the **expectation** of a function, such as $E[g(x)] = \sum_{x \in \mathcal{X}} g(x)p(x)$. Expectations are easiest to think about in terms of probability distributions over discrete events:

- If it is sunny, Lucia will eat three ice creams.
- If it is rainy, she will eat only one ice cream.
- There's a 80% chance it will be sunny.
- The expected number of ice creams she will eat is $0.8 \times 3 + 0.2 \times 1 = 2.6$.

If the random variable X is continuous, the expectation is an integral:

$$E[g(x)] = \int_{\mathcal{X}} g(x)p(x)dx \quad [A.30]$$

For example, a fast food restaurant in Quebec has a special offer for cold days: they give a 1% discount on poutine for every degree below zero. Assuming a thermometer with infinite precision, the expected price would be an integral over all possible temperatures,

$$E[\text{price}(x)] = \int_{\mathcal{X}} \min(1, 1+x) \times \text{original-price} \times p(x)dx. \quad [A.31]$$

11039 **A.6 Modeling and estimation**

11040 **Probabilistic models** provide a principled way to reason about random events and ran-
11041 dom variables. Let's consider the coin toss example. Each toss can be modeled as a ran-
11042 dom event, with probability θ of the event H , and probability $1 - \theta$ of the complementary
11043 event T . If we write a random variable X as the total number of heads on three coin
11044 flips, then the distribution of X depends on θ . In this case, X is distributed as a **binomial**
11045 **random variable**, meaning that it is drawn from a binomial distribution, with **parameters**
11046 $(\theta, N = 3)$. This is written,

$$X \sim \text{Binomial}(\theta, N = 3). \quad [\text{A.32}]$$

11047 The properties of the binomial distribution enable us to make statements about the X ,
11048 such as its expected value and the likelihood that its value will fall within some interval.

Now suppose that θ is unknown, but we have run an experiment, in which we ex-
 ecuted N trials, and obtained x heads. We can **estimate** θ by the principle of **maximum**
likelihood:

$$\hat{\theta} = \operatorname{argmax}_{\theta} p_X(x; \theta, N). \quad [\text{A.33}]$$

This says that the estimate $\hat{\theta}$ should be the value that maximizes the likelihood of the
 data. The semicolon indicates that θ and N are parameters of the probability function.
 The likelihood $p_X(x; \theta, N)$ can be computed from the binomial distribution,

$$p_X(x; \theta, N) = \frac{N!}{x!(N-x)!} \theta^x (1-\theta)^{N-x}. \quad [\text{A.34}]$$

11049 This likelihood is proportional to the product of the probability of individual out-
11050 comes: for example, the sequence T, H, H, T, H would have probability $\theta^3(1-\theta)^2$. The
11051 term $\frac{N!}{x!(N-x)!}$ arises from the many possible orderings by which we could obtain x heads
11052 on N trials. This term does not depend on θ , so it can be ignored during estimation.

In practice, we maximize the log-likelihood, which is a monotonic function of the like-
 lihood. Under the binomial distribution, the log-likelihood is a **convex** function of θ (see

§ 2.4), so it can be maximized by taking the derivative and setting it equal to zero.

$$\ell(\theta) = x \log \theta + (N - x) \log(1 - \theta) \quad [\text{A.35}]$$

$$\frac{\partial \ell(\theta)}{\partial \theta} = \frac{x}{\theta} - \frac{N - x}{1 - \theta} \quad [\text{A.36}]$$

$$\frac{N - x}{1 - \theta} = \frac{x}{\theta} \quad [\text{A.37}]$$

$$\frac{N - x}{x} = \frac{1 - \theta}{\theta} \quad [\text{A.38}]$$

$$\frac{N}{x} - 1 = \frac{1}{\theta} - 1 \quad [\text{A.39}]$$

$$\hat{\theta} = \frac{x}{N}. \quad [\text{A.40}]$$

11053 In this case, the maximum likelihood estimate is equal to $\frac{x}{N}$, the fraction of trials that
 11054 came up heads. This intuitive solution is also known as the **relative frequency estimate**,
 11055 since it is equal to the relative frequency of the outcome.

Is maximum likelihood estimation always the right choice? Suppose you conduct one trial, and get heads. Would you conclude that $\theta = 1$, meaning that the coin is guaranteed to come up heads? If not, then you must have some **prior expectation** about θ . To incorporate this prior information, we can treat θ as a random variable, and use Bayes' rule:

$$p(\theta | x; N) = \frac{p(x | \theta) \times p(\theta)}{p(x)} \quad [\text{A.41}]$$

$$\propto p(x | \theta) \times p(\theta) \quad [\text{A.42}]$$

$$\hat{\theta} = \operatorname{argmax}_{\theta} p(x | \theta) \times p(\theta). \quad [\text{A.43}]$$

11056 This is the **maximum a posteriori** (MAP) estimate. Given a form for $p(\theta)$, you can de-
 11057 rive the MAP estimate using the same approach that was used to derive the maximum
 11058 likelihood estimate.

11059 Additional resources

11060 A good introduction to probability theory is offered by Manning and Schütze (1999),
 11061 which helped to motivate this section. For more detail, Sharon Goldwater provides an-
 11062 other useful reference, <http://homepages.inf.ed.ac.uk/sgwater/teaching/general/probability.pdf>. A historical and philosophical perspective on probability is offered
 11063 by Diaconis and Skyrms (2017).

11065 **Appendix B**

11066 **Numerical optimization**

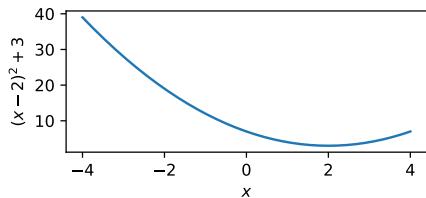
11067 Unconstrained numerical optimization involves solving problems of the form,

$$\min_{\mathbf{x} \in \mathbb{R}^D} f(\mathbf{x}), \quad [\text{B.1}]$$

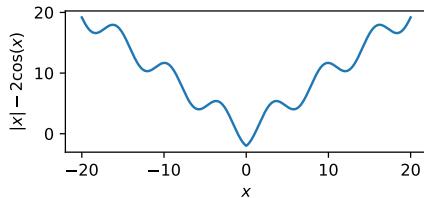
11068 where $\mathbf{x} \in \mathbb{R}^D$ is a vector of D real numbers.

11069 Differentiation is fundamental to numerical optimization. Suppose that at some \mathbf{x}^* ,
11070 every partial derivative of f is equal to 0: formally, $\frac{\partial f}{\partial x_i} \Big|_{\mathbf{x}^*} = 0$. Then \mathbf{x}^* is said to be a
11071 **critical point** of f . If f is a **convex** function (defined in § 2.4), then the value of $f(\mathbf{x}^*)$ is
11072 equal to the global minimum of f iff \mathbf{x}^* is a critical point of f .

As an example, consider the convex function $f(x) = (x - 2)^2 + 3$, shown in Figure B.1a. The derivative is $\frac{\partial f}{\partial x} = 2x - 4$. A unique minimum can be obtained by setting the derivative equal to zero and solving for x , obtaining $x^* = 2$. Now consider the multivariate convex function $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x} - [2, 1]^\top\|^2$, where $\|\mathbf{x}\|^2$ is the squared Euclidean norm. The partial



(a) The function $f(x) = (x - 2)^2 + 3$



(b) The function $f(x) = |x| - 2\cos(x)$

Figure B.1: Two functions with unique global minima

derivatives are,

$$\frac{\partial d}{\partial x_1} = x_1 - 2 \quad [B.2]$$

$$\frac{\partial d}{\partial x_2} = x_2 - 1 \quad [B.3]$$

11073 The unique minimum is $\mathbf{x}^* = [2, 1]^\top$.

11074 For non-convex functions, critical points are not necessarily global minima. A **local**
 11075 **minimum** \mathbf{x}^* is a point at which the function takes a smaller value than at all nearby
 11076 neighbors: formally, \mathbf{x}^* is a local minimum if there is some positive ϵ such that $f(\mathbf{x}^*) \leq$
 11077 $f(\mathbf{x})$ for all \mathbf{x} within distance ϵ of \mathbf{x}^* . Figure B.1b shows the function $f(x) = |x| - 2 \cos(x)$,
 11078 which has many local minima, as well as a unique global minimum at $x = 0$. A critical
 11079 point may also be the local or global maximum of the function; it may be a **saddle point**,
 11080 which is a minimum with respect to at least one coordinate, and a maximum with respect
 11081 to at least one other coordinate; it may be an **inflection point**, which is neither a minimum
 11082 nor maximum. When available, the second derivative of f can help to distinguish these
 11083 cases.

11084 B.1 Gradient descent

For many convex functions, it is not possible to solve for \mathbf{x}^* in closed form. In gradient descent, we compute a series of solutions, $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \dots$ by taking steps along the local gradient $\nabla_{\mathbf{x}^{(t)}} f$, which is the vector of partial derivatives of the function f , evaluated at the point $\mathbf{x}^{(t)}$. Each solution $\mathbf{x}^{(t+1)}$ is computed,

$$\mathbf{x}^{(t+1)} \leftarrow \mathbf{x}^{(t)} - \eta^{(t)} \nabla_{\mathbf{x}^{(t)}} f. \quad [B.4]$$

11085 where $\eta^{(t)} > 0$ is a **step size**. If the step size is chosen appropriately, this procedure will
 11086 find the global minimum of a differentiable convex function. For non-convex functions,
 11087 gradient descent will find a local minimum. The extension to non-differentiable convex
 11088 functions is discussed in § 2.4.

11089 B.2 Constrained optimization

Optimization must often be performed under constraints: for example, when optimizing the parameters of a probability distribution, the probabilities of all events must sum to one. Constrained optimization problems can be written,

$$\min_{\mathbf{x}} f(\mathbf{x}) \quad [B.5]$$

$$\text{s.t. } g_c(\mathbf{x}) \leq 0, \quad \forall c = 1, 2, \dots, C \quad [B.6]$$

where each $g_c(\mathbf{x})$ is a scalar function of \mathbf{x} . For example, suppose that \mathbf{x} must be non-negative, and that its sum cannot exceed a budget b . Then there are $D + 1$ inequality constraints,

$$g_i(\mathbf{x}) = -x_i, \quad \forall i = 1, 2, \dots, D \quad [\text{B.7}]$$

$$g_{D+1}(\mathbf{x}) = -b + \sum_{i=1}^D x_i. \quad [\text{B.8}]$$

Inequality constraints can be combined with the original objective function f by forming a **Lagrangian**,

$$L(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \sum_{c=1}^C \lambda_c g_c(\mathbf{x}), \quad [\text{B.9}]$$

where λ_c is a **Lagrange multiplier**. For any Lagrangian, there is a corresponding dual form, which is a function of $\boldsymbol{\lambda}$:

$$D(\boldsymbol{\lambda}) = \min_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}). \quad [\text{B.10}]$$

The Lagrangian L can be referred to as the **primal form**.

B.3 Example: Passive-aggressive online learning

Sometimes it is possible to solve a constrained optimization problem by manipulating the Lagrangian. One example is maximum-likelihood estimation of a Naïve Bayes probability model, as described in § 2.2.3. In that case, it is unnecessary to explicitly compute the Lagrange multiplier. Another example is illustrated by the **passive-aggressive** algorithm for online learning (Crammer et al., 2006). This algorithm is similar to the perceptron, but the goal at each step is to make the most conservative update that gives zero margin loss on the current example.¹ Each update can be formulated as a constrained optimization over the weights $\boldsymbol{\theta}$:

$$\min_{\boldsymbol{\theta}} \frac{1}{2} \|\boldsymbol{\theta} - \boldsymbol{\theta}^{(i-1)}\|^2 \quad [\text{B.11}]$$

$$\text{s.t. } \ell^{(i)}(\boldsymbol{\theta}) = 0 \quad [\text{B.12}]$$

where $\boldsymbol{\theta}^{(i-1)}$ is the previous set of weights, and $\ell^{(i)}(\boldsymbol{\theta})$ is the margin loss on instance i . As in § 2.4.1, this loss is defined as,

$$\ell^{(i)}(\boldsymbol{\theta}) = 1 - \boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{x}^{(i)}, y^{(i)}) + \max_{y \neq y^{(i)}} \boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{x}^{(i)}, y). \quad [\text{B.13}]$$

¹This is the basis for the name of the algorithm: it is passive when the loss is zero, but it aggressively moves to make the loss zero when necessary.

When the margin loss is zero for $\theta^{(i-1)}$, the optimal solution is $\theta^* = \theta^{(i-1)}$, so we will focus on the case where $\ell^{(i)}(\theta^{(i-1)}) > 0$. The Lagrangian for this problem is,

$$L(\theta, \lambda) = \frac{1}{2} \|\theta - \theta^{(i-1)}\|^2 + \lambda \ell^{(i)}(\theta), \quad [\text{B.14}]$$

Holding λ constant, we can solve for θ by differentiating,

$$\nabla_{\theta} L = \theta - \theta^{(i-1)} + \lambda \frac{\partial}{\partial \theta} \ell^{(i)}(\theta) \quad [\text{B.15}]$$

$$\theta^* = \theta^{(i-1)} + \lambda \delta, \quad [\text{B.16}]$$

where $\delta = f(x^{(i)}, y^{(i)}) - f(x^{(i)}, \hat{y})$ and $\hat{y} = \operatorname{argmax}_{y \neq y^{(i)}} \theta \cdot f(x^{(i)}, y)$.

The Lagrange multiplier λ acts as the learning rate in a perceptron-style update to θ . We can solve for λ by plugging θ^* back into the Lagrangian, obtaining the dual function,

$$D(\lambda) = \frac{1}{2} \|\theta^{(i-1)} + \lambda \delta - \theta^{(i-1)}\|^2 + \lambda(1 - (\theta^{(i-1)} + \lambda \delta) \cdot \delta) \quad [\text{B.17}]$$

$$= \frac{\lambda^2}{2} \|\delta\|^2 - \lambda^2 \|\delta\|^2 + \lambda(1 - \theta^{(i-1)} \cdot \delta) \quad [\text{B.18}]$$

$$= -\frac{\lambda^2}{2} \|\delta\|^2 + \lambda \ell^{(i)}(\theta^{(i-1)}). \quad [\text{B.19}]$$

Differentiating and solving for λ ,

$$\frac{\partial D}{\partial \lambda} = -\lambda \|\delta\|^2 + \ell^{(i)}(\theta^{(i-1)}) \quad [\text{B.20}]$$

$$\lambda^* = \frac{\ell^{(i)}(\theta^{(i-1)})}{\|\delta\|^2}. \quad [\text{B.21}]$$

The complete update equation is therefore:

$$\theta^* = \theta^{(i-1)} + \frac{\ell^{(i)}(\theta^{(i-1)})}{\|f(x^{(i)}, y^{(i)}) - f(x^{(i)}, \hat{y})\|^2} (f(x^{(i)}, y^{(i)}) - f(x^{(i)}, \hat{y})). \quad [\text{B.22}]$$

This learning rate makes intuitive sense. The numerator grows with the loss; the denominator grows with the norm of the difference between the feature vectors associated with the correct and predicted label. If this norm is large, then the step with respect to each feature should be small, and vice versa.

11106

Bibliography

- 11107 Abadi, M., A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis,
11108 J. Dean, M. Devin, S. Ghemawat, I. J. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia,
11109 R. Józefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore,
11110 D. G. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. A.
11111 Tucker, V. Vanhoucke, V. Vasudevan, F. B. Viégas, O. Vinyals, P. Warden, M. Watten-
11112 berg, M. Wicke, Y. Yu, and X. Zheng (2016). Tensorflow: Large-scale machine learning
11113 on heterogeneous distributed systems. *CoRR abs/1603.04467*.
- 11114 Abend, O. and A. Rappoport (2017). The state of the art in semantic representation. In
11115 *Proceedings of the Association for Computational Linguistics (ACL)*.
- 11116 Abney, S., R. E. Schapire, and Y. Singer (1999). Boosting applied to tagging and PP attach-
11117 ment. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp.
11118 132–134.
- 11119 Abney, S. P. (1987). *The English noun phrase in its sentential aspect*. Ph. D. thesis, Mas-
11120 sachusetts Institute of Technology.
- 11121 Abney, S. P. and M. Johnson (1991). Memory requirements and local ambiguities of pars-
11122 ing strategies. *Journal of Psycholinguistic Research* 20(3), 233–250.
- 11123 Adafre, S. F. and M. De Rijke (2006). Finding similar sentences across multiple languages
11124 in wikipedia. In *Proceedings of the Workshop on NEW TEXT Wikis and blogs and other*
11125 *dynamic text sources*.
- 11126 Ahn, D. (2006). The stages of event extraction. In *Proceedings of the Workshop on Annotating*
11127 *and Reasoning about Time and Events*, pp. 1–8. Association for Computational Linguistics.
- 11128 Aho, A. V., M. S. Lam, R. Sethi, and J. D. Ullman (2006). *Compilers: Principles, Techniques,*
11129 *& Tools* (2nd ed.). Addison-Wesley Publishing Company.
- 11130 Aikhenvald, A. Y. (2004). *Evidentiality*. Oxford University Press.

- 11131 Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on
11132 Automatic Control* 19(6), 716–723.
- 11133 Akmajian, A., R. A. Demers, A. K. Farmer, and R. M. Harnish (2010). *Linguistics: An
11134 introduction to language and communication* (Sixth ed.). Cambridge, MA: MIT press.
- 11135 Alfano, M., D. Hovy, M. Mitchell, and M. Strube (2018). Proceedings of the second acl
11136 workshop on ethics in natural language processing. In *Proceedings of the Second ACL
11137 Workshop on Ethics in Natural Language Processing*. Association for Computational Lin-
11138 guistics.
- 11139 Alfau, F. (1999). *Chromos*. Dalkey Archive Press.
- 11140 Allauzen, C., M. Riley, J. Schalkwyk, W. Skut, and M. Mohri (2007). OpenFst: A gen-
11141 eral and efficient weighted finite-state transducer library. In *International Conference on
11142 Implementation and Application of Automata*, pp. 11–23. Springer.
- 11143 Allen, J. F. (1984). Towards a general theory of action and time. *Artificial intelligence* 23(2),
11144 123–154.
- 11145 Allen, J. F., B. W. Miller, E. K. Ringger, and T. Sikorski (1996). A robust system for natural
11146 spoken dialogue. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp.
11147 62–70.
- 11148 Allen, J. F., L. K. Schubert, G. Ferguson, P. Heeman, C. H. Hwang, T. Kato, M. Light,
11149 N. Martin, B. Miller, M. Poesio, and D. Traum (1995). The TRAINS project: A case
11150 study in building a conversational planning agent. *Journal of Experimental & Theoretical
11151 Artificial Intelligence* 7(1), 7–48.
- 11152 Alm, C. O., D. Roth, and R. Sproat (2005). Emotions from text: machine learning for
11153 text-based emotion prediction. In *Proceedings of Empirical Methods for Natural Language
11154 Processing (EMNLP)*, pp. 579–586.
- 11155 Aluísio, S., J. Pelizzoni, A. Marchi, L. de Oliveira, R. Manenti, and V. Marquiafável (2003).
11156 An account of the challenge of tagging a reference corpus for Brazilian Portuguese.
11157 *Computational Processing of the Portuguese Language*, 194–194.
- 11158 Anand, P., M. Walker, R. Abbott, J. E. Fox Tree, R. Bowman, and M. Minor (2011). Cats rule
11159 and dogs drool!: Classifying stance in online debate. In *Proceedings of the 2nd Workshop
11160 on Computational Approaches to Subjectivity and Sentiment Analysis*, pp. 1–9. Association
11161 for Computational Linguistics.
- 11162 Anandkumar, A. and R. Ge (2016). Efficient approaches for escaping higher order saddle
11163 points in non-convex optimization. In *Proceedings of the Conference On Learning Theory
11164 (COLT)*, pp. 81–102.

- 11165 Anandkumar, A., R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky (2014). Tensor decompo-
11166 sitions for learning latent variable models. *The Journal of Machine Learning Research* 15(1),
11167 2773–2832.
- 11168 Ando, R. K. and T. Zhang (2005). A framework for learning predictive structures from
11169 multiple tasks and unlabeled data. *The Journal of Machine Learning Research* 6, 1817–
11170 1853.
- 11171 Andor, D., C. Alberti, D. Weiss, A. Severyn, A. Presta, K. Ganchev, S. Petrov, and
11172 M. Collins (2016). Globally normalized transition-based neural networks. In *Proceedings*
11173 of the Association for Computational Linguistics (ACL), pp. 2442–2452.
- 11174 Angeli, G., P. Liang, and D. Klein (2010). A simple domain-independent probabilistic ap-
11175 proach to generation. In *Proceedings of Empirical Methods for Natural Language Processing*
11176 (EMNLP), pp. 502–512.
- 11177 Antol, S., A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh
11178 (2015). Vqa: Visual question answering. In *Proceedings of the International Conference on*
11179 *Computer Vision (ICCV)*, pp. 2425–2433.
- 11180 Aronoff, M. (1976). *Word formation in generative grammar*. MIT Press.
- 11181 Arora, S. and B. Barak (2009). *Computational complexity: a modern approach*. Cambridge
11182 University Press.
- 11183 Arora, S., R. Ge, Y. Halpern, D. Mimno, A. Moitra, D. Sontag, Y. Wu, and M. Zhu (2013).
11184 A practical algorithm for topic modeling with provable guarantees. In *Proceedings of the*
11185 *International Conference on Machine Learning (ICML)*, pp. 280–288.
- 11186 Arora, S., Y. Li, Y. Liang, T. Ma, and A. Risteski (2018). Linear algebraic structure of word
11187 senses, with applications to polysemy. *Transactions of the Association of Computational*
11188 *Linguistics* 6, 483–495.
- 11189 Artstein, R. and M. Poesio (2008). Inter-coder agreement for computational linguistics.
11190 *Computational Linguistics* 34(4), 555–596.
- 11191 Artzi, Y. and L. Zettlemoyer (2013). Weakly supervised learning of semantic parsers for
11192 mapping instructions to actions. *Transactions of the Association for Computational Linguis-*
11193 *tics* 1, 49–62.
- 11194 Attardi, G. (2006). Experiments with a multilanguage non-projective dependency parser.
11195 In *Proceedings of the Conference on Natural Language Learning (CoNLL)*, pp. 166–170.
- 11196 Auer, P. (2013). *Code-switching in conversation: Language, interaction and identity*. Routledge.

- 11197 Auer, S., C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives (2007). Dbpedia: A
11198 nucleus for a web of open data. *The semantic web*, 722–735.
- 11199 Austin, J. L. (1962). *How to do things with words*. Oxford University Press.
- 11200 Aw, A., M. Zhang, J. Xiao, and J. Su (2006). A phrase-based statistical model for SMS text
11201 normalization. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp.
11202 33–40.
- 11203 Ba, J. L., J. R. Kiros, and G. E. Hinton (2016). Layer normalization. *arXiv preprint*
11204 *arXiv:1607.06450*.
- 11205 Bagga, A. and B. Baldwin (1998a). Algorithms for scoring coreference chains. In *Proceed-
11206 ings of the Language Resources and Evaluation Conference*, pp. 563–566.
- 11207 Bagga, A. and B. Baldwin (1998b). Entity-based cross-document coreferencing using the
11208 vector space model. In *Proceedings of the International Conference on Computational Lin-
11209 guistics (COLING)*, pp. 79–85.
- 11210 Bahdanau, D., K. Cho, and Y. Bengio (2014). Neural machine translation by jointly learn-
11211 ing to align and translate. In *Neural Information Processing Systems (NIPS)*.
- 11212 Baldwin, T. and S. N. Kim (2010). Multiword expressions. In *Handbook of natural language
11213 processing*, Volume 2, pp. 267–292. Boca Raton, USA: CRC Press.
- 11214 Balle, B., A. Quattoni, and X. Carreras (2011). A spectral learning algorithm for finite state
11215 transducers. In *Proceedings of the European Conference on Machine Learning and Principles
11216 and Practice of Knowledge Discovery in Databases (ECML)*, pp. 156–171.
- 11217 Banarescu, L., C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight,
11218 P. Koehn, M. Palmer, and N. Schneider (2013, August). Abstract meaning represen-
11219 tation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability
11220 with Discourse*, Sofia, Bulgaria, pp. 178–186. Association for Computational
11221 Linguistics.
- 11222 Banko, M., M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni (2007). Open
11223 information extraction from the web. In *Proceedings of the International Joint Conference
11224 on Artificial Intelligence (IJCAI)*, pp. 2670–2676.
- 11225 Bansal, N., A. Blum, and S. Chawla (2004). Correlation clustering. *Machine Learning* 56(1–
11226 3), 89–113.
- 11227 Barber, D. (2012). *Bayesian reasoning and machine learning*. Cambridge University Press.

- 11228 Barman, U., A. Das, J. Wagner, and J. Foster (2014). Code mixing: A challenge for language
11229 identification in the language of social media. In *Proceedings of the First Workshop on*
11230 *Computational Approaches to Code Switching*, pp. 13–23. Association for Computational
11231 Linguistics.
- 11232 Baron, A. and P. Rayson (2008). Vard2: A tool for dealing with spelling variation in his-
11233 torical corpora. In *Postgraduate conference in corpus linguistics*.
- 11234 Baroni, M., R. Bernardi, and R. Zamparelli (2014). Frege in space: A program for compo-
11235 sitional distributional semantics. *Linguistic Issues in Language Technologies*.
- 11236 Barzilay, R. and M. Lapata (2008). Modeling local coherence: An Entity-Based approach.
11237 *Computational Linguistics* 34(1), 1–34.
- 11238 Barzilay, R. and K. R. McKeown (2005). Sentence fusion for multidocument news summa-
11239 rization. *Computational Linguistics* 31(3), 297–328.
- 11240 Beesley, K. R. and L. Karttunen (2003). *Finite-state morphology*. Stanford, CA: Center for
11241 the Study of Language and Information.
- 11242 Bejan, C. A. and S. Harabagiu (2014). Unsupervised event coreference resolution. *Compu-*
11243 *tational Linguistics* 40(2), 311–347.
- 11244 Bell, E. T. (1934). Exponential numbers. *The American Mathematical Monthly* 41(7), 411–419.
- 11245 Bender, E. M. (2013). *Linguistic Fundamentals for Natural Language Processing: 100 Essentials*
11246 *from Morphology and Syntax*, Volume 6 of *Synthesis Lectures on Human Language Technolo-*
11247 *gies*. Morgan & Claypool Publishers.
- 11248 Bengio, S., O. Vinyals, N. Jaitly, and N. Shazeer (2015). Scheduled sampling for sequence
11249 prediction with recurrent neural networks. In *Neural Information Processing Systems*
11250 (*NIPS*), pp. 1171–1179.
- 11251 Bengio, Y., R. Ducharme, P. Vincent, and C. Janvin (2003). A neural probabilistic language
11252 model. *The Journal of Machine Learning Research* 3, 1137–1155.
- 11253 Bengio, Y., P. Simard, and P. Frasconi (1994). Learning long-term dependencies with gra-
11254 dient descent is difficult. *IEEE Transactions on Neural Networks* 5(2), 157–166.
- 11255 Bengtson, E. and D. Roth (2008). Understanding the value of features for coreference
11256 resolution. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*,
11257 pp. 294–303.
- 11258 Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and
11259 powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B*
11260 (*Methodological*), 289–300.

- 11261 Berant, J., A. Chou, R. Frostig, and P. Liang (2013). Semantic parsing on freebase from
11262 question-answer pairs. In *Proceedings of Empirical Methods for Natural Language Processing*
11263 (*EMNLP*), pp. 1533–1544.
- 11264 Berant, J., V. Srikumar, P.-C. Chen, A. Vander Linden, B. Harding, B. Huang, P. Clark, and
11265 C. D. Manning (2014). Modeling biological processes for reading comprehension. In
11266 *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.
- 11267 Berg-Kirkpatrick, T., A. Bouchard-Côté, J. DeNero, and D. Klein (2010). Painless unsuper-
11268 vised learning with features. In *Proceedings of the North American Chapter of the Associa-*
11269 *tion for Computational Linguistics (NAACL)*, pp. 582–590.
- 11270 Berg-Kirkpatrick, T., D. Burkett, and D. Klein (2012). An empirical investigation of sta-
11271 tistical significance in NLP. In *Proceedings of Empirical Methods for Natural Language*
11272 *Processing (EMNLP)*, pp. 995–1005.
- 11273 Berger, A. L., V. J. D. Pietra, and S. A. D. Pietra (1996). A maximum entropy approach to
11274 natural language processing. *Computational linguistics* 22(1), 39–71.
- 11275 Bergsma, S., D. Lin, and R. Goebel (2008). Distributional identification of non-referential
11276 pronouns. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 10–18.
- 11277 Bernardi, R., R. Cakici, D. Elliott, A. Erdem, E. Erdem, N. Ikizler-Cinbis, F. Keller, A. Mus-
11278 cat, and B. Plank (2016). Automatic description generation from images: A survey of
11279 models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research* 55,
11280 409–442.
- 11281 Bertsekas, D. P. (2012). Incremental gradient, subgradient, and proximal methods for con-
11282 vex optimization: A survey. In S. Sra, S. Nowozin, and S. J. Wright (Eds.), *Optimization*
11283 for machine learning. MIT Press.
- 11284 Bhatia, P., R. Guthrie, and J. Eisenstein (2016). Morphological priors for probabilistic neu-
11285 ral word embeddings. In *Proceedings of Empirical Methods for Natural Language Processing*
11286 (*EMNLP*).
- 11287 Bhatia, P., Y. Ji, and J. Eisenstein (2015). Better document-level sentiment analysis from
11288 first discourse parsing. In *Proceedings of Empirical Methods for Natural Language Processing*
11289 (*EMNLP*).
- 11290 Biber, D. (1991). *Variation across speech and writing*. Cambridge University Press.
- 11291 Bird, S., E. Klein, and E. Loper (2009). *Natural language processing with Python*. California:
11292 O'Reilly Media.
- 11293 Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

- 11294 Björkelund, A. and P. Nugues (2011). Exploring lexicalized features for coreference reso-
11295 lution. In *Proceedings of the Conference on Natural Language Learning (CoNLL)*, pp. 45–50.
- 11296 Blackburn, P. and J. Bos (2005). *Representation and inference for natural language: A first*
11297 *course in computational semantics*. CSLI.
- 11298 Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM* 55(4), 77–84.
- 11299 Blei, D. M. (2014). Build, compute, critique, repeat: Data analysis with latent variable
11300 models. *Annual Review of Statistics and Its Application* 1, 203–232.
- 11301 Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent dirichlet allocation. *The Journal of*
11302 *Machine Learning Research* 3, 993–1022.
- 11303 Blitzer, J., M. Dredze, and F. Pereira (2007). Biographies, bollywood, boom-boxes and
11304 blenders: Domain adaptation for sentiment classification. In *Proceedings of the Associa-*
11305 *tion for Computational Linguistics (ACL)*, pp. 440–447.
- 11306 Blum, A. and T. Mitchell (1998). Combining labeled and unlabeled data with co-training.
11307 In *Proceedings of the Conference On Learning Theory (COLT)*, pp. 92–100.
- 11308 Bobrow, D. G., R. M. Kaplan, M. Kay, D. A. Norman, H. Thompson, and T. Winograd
11309 (1977). Gus, a frame-driven dialog system. *Artificial intelligence* 8(2), 155–173.
- 11310 Bochnet, B. (2010). Very high accuracy and fast dependency parsing is not a contradiction.
11311 In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pp.
11312 89–97.
- 11313 Boitet, C. (1988). Pros and cons of the pivot and transfer approaches in multilingual ma-
11314 chine translation. *Readings in machine translation*, 273–279.
- 11315 Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov (2017). Enriching word vectors with
11316 subword information. *Transactions of the Association for Computational Linguistics* 5, 135–
11317 146.
- 11318 Bollacker, K., C. Evans, P. Paritosh, T. Sturge, and J. Taylor (2008). Freebase: a collabora-
11319 tively created graph database for structuring human knowledge. In *Proceedings of the*
11320 *ACM International Conference on Management of Data (SIGMOD)*, pp. 1247–1250. AcM.
- 11321 Bolukbasi, T., K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai (2016). Man is to
11322 computer programmer as woman is to homemaker? debiasing word embeddings. In *Neural Information Processing Systems (NIPS)*, pp. 4349–4357.
- 11324 Bordes, A., N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko (2013). Translating
11325 embeddings for modeling multi-relational data. In *Neural Information Processing Systems*
11326 (*NIPS*), pp. 2787–2795.

- 11327 Bordes, A., J. Weston, R. Collobert, Y. Bengio, et al. (2011). Learning structured embed-
 11328 dings of knowledge bases. In *Proceedings of the National Conference on Artificial Intelligence
 11329 (AAAI)*, pp. 301–306.
- 11330 Borges, J. L. (1993). *Other Inquisitions 1937–1952*. University of Texas Press. Translated by
 11331 Ruth L. C. Simms.
- 11332 Botha, J. A. and P. Blunsom (2014). Compositional morphology for word representations
 11333 and language modelling. In *Proceedings of the International Conference on Machine Learn-
 11334 ing (ICML)*.
- 11335 Bottou, L. (2012). Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade*,
 11336 pp. 421–436. Springer.
- 11337 Bottou, L., F. E. Curtis, and J. Nocedal (2016). Optimization methods for large-scale ma-
 11338 chine learning. *arXiv preprint arXiv:1606.04838*.
- 11339 Bowman, S. R., L. Vilnis, O. Vinyals, A. Dai, R. Jozefowicz, and S. Bengio (2016). Gen-
 11340 erating sentences from a continuous space. In *Proceedings of the Conference on Natural
 11341 Language Learning (CoNLL)*, pp. 10–21.
- 11342 boyd, d. and K. Crawford (2012). Critical questions for big data. *Information, Communica-
 11343 tion & Society* 15(5), 662–679.
- 11344 Boyd, S. and L. Vandenberghe (2004). *Convex Optimization*. New York: Cambridge Uni-
 11345 versity Press.
- 11346 Boydston, A. E. (2013). *Making the news: Politics, the media, and agenda setting*. University
 11347 of Chicago Press.
- 11348 Branavan, S., H. Chen, J. Eisenstein, and R. Barzilay (2009). Learning document-level
 11349 semantic properties from free-text annotations. *Journal of Artificial Intelligence Re-
 11350 search* 34(2), 569–603.
- 11351 Branavan, S. R., H. Chen, L. S. Zettlemoyer, and R. Barzilay (2009). Reinforcement learning
 11352 for mapping instructions to actions. In *Proceedings of the Association for Computational
 11353 Linguistics (ACL)*, pp. 82–90.
- 11354 Brants, T. and A. Franz (2006). The Google 1T 5-gram corpus. LDC2006T13.
- 11355 Braud, C., O. Lacroix, and A. Søgaard (2017). Does syntax help discourse segmenta-
 11356 tion? not so much. In *Proceedings of Empirical Methods for Natural Language Processing
 11357 (EMNLP)*, pp. 2432–2442.
- 11358 Brennan, S. E., M. W. Friedman, and C. J. Pollard (1987). A centering approach to pro-
 11359 nouns. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 155–162.

- 11360 Briscoe, T. (2011). Introduction to formal semantics for natural language.
11361 <https://www.cl.cam.ac.uk/teaching/1011/L107/semantics.pdf>.
- 11362 Brown, P. F., J. Cocke, S. A. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer,
11363 and P. S. Roossin (1990). A statistical approach to machine translation. *Computational
11364 linguistics* 16(2), 79–85.
- 11365 Brown, P. F., P. V. Desouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai (1992). Class-based
11366 n-gram models of natural language. *Computational linguistics* 18(4), 467–479.
- 11367 Brown, P. F., V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer (1993). The mathematics
11368 of statistical machine translation: Parameter estimation. *Computational linguistics* 19(2),
11369 263–311.
- 11370 Brun, C. and C. Roux (2014). Décomposition des “hash tags” pour l’amélioration de la
11371 classification en polarité des “tweets”. *Proceedings of Traitement Automatique des Langues
11372 Naturelles*, 473–478.
- 11373 Bruni, E., N.-K. Tran, and M. Baroni (2014). Multimodal distributional semantics. *Journal
11374 of Artificial Intelligence Research* 49(2014), 1–47.
- 11375 Brutzkus, A., A. Globerson, E. Malach, and S. Shalev-Shwartz (2018). Sgd learns over-
11376 parameterized networks that provably generalize on linearly separable data. In *Pro-
11377 ceedings of the International Conference on Learning Representations (ICLR)*.
- 11378 Bullinaria, J. A. and J. P. Levy (2007). Extracting semantic representations from word co-
11379 occurrence statistics: A computational study. *Behavior research methods* 39(3), 510–526.
- 11380 Bunescu, R. C. and R. J. Mooney (2005). A shortest path dependency kernel for relation
11381 extraction. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*,
11382 pp. 724–731.
- 11383 Bunescu, R. C. and M. Pasca (2006). Using encyclopedic knowledge for named entity
11384 disambiguation. In *Proceedings of the European Chapter of the Association for Computational
11385 Linguistics (EACL)*, pp. 9–16.
- 11386 Burstein, J., D. Marcu, and K. Knight (2003). Finding the WRITE stuff: Automatic identi-
11387 fication of discourse structure in student essays. *IEEE Intelligent Systems* 18(1), 32–39.
- 11388 Burstein, J., J. Tetreault, and S. Andreyev (2010). Using entity-based features to model
11389 coherence in student essays. In *Human language technologies: The 2010 annual conference
11390 of the North American chapter of the Association for Computational Linguistics*, pp. 681–684.
11391 Association for Computational Linguistics.

- 11392 Burstein, J., J. Tetreault, and M. Chodorow (2013). Holistic discourse coherence annotation
 11393 for noisy essay writing. *Dialogue & Discourse* 4(2), 34–52.
- 11394 Cai, Q. and A. Yates (2013). Large-scale semantic parsing via schema matching and lexicon
 11395 extension. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 423–
 11396 433.
- 11397 Caliskan, A., J. J. Bryson, and A. Narayanan (2017). Semantics derived automatically from
 11398 language corpora contain human-like biases. *Science* 356(6334), 183–186.
- 11399 Canny, J. (1987). A computational approach to edge detection. In *Readings in Computer
 11400 Vision*, pp. 184–203. Elsevier.
- 11401 Cappé, O. and E. Moulines (2009). On-line expectation–maximization algorithm for latent
 11402 data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71(3),
 11403 593–613.
- 11404 Carbonell, J. and J. Goldstein (1998). The use of MMR, diversity-based reranking for re-
 11405 ordering documents and producing summaries. In *Proceedings of ACM SIGIR conference
 11406 on Research and development in information retrieval*.
- 11407 Carbonell, J. R. (1970). Mixed-initiative man-computer instructional dialogues. Technical
 11408 report, Bolt Beranek and Newman.
- 11409 Cardie, C. and K. Wagstaff (1999). Noun phrase coreference as clustering. In *Proceedings
 11410 of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 82–89.
- 11411 Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistic. *Com-
 11412 putational linguistics* 22(2), 249–254.
- 11413 Carletta, J. (2007). Unleashing the killer corpus: experiences in creating the multi-
 11414 everything ami meeting corpus. *Language Resources and Evaluation* 41(2), 181–190.
- 11415 Carlson, L. and D. Marcu (2001). Discourse tagging reference manual. Technical Report
 11416 ISI-TR-545, Information Sciences Institute.
- 11417 Carlson, L., M. E. Okurowski, and D. Marcu (2002). RST discourse treebank. Linguistic
 11418 Data Consortium, University of Pennsylvania.
- 11419 Carpenter, B. (1997). *Type-logical semantics*. Cambridge, MA: MIT Press.
- 11420 Carreras, X., M. Collins, and T. Koo (2008). Tag, dynamic programming, and the percep-
 11421 tron for efficient, feature-rich parsing. In *Proceedings of the Conference on Natural Language
 11422 Learning (CoNLL)*, pp. 9–16.

- 11423 Carreras, X. and L. Màrquez (2005). Introduction to the conll-2005 shared task: Semantic
11424 role labeling. In *Proceedings of the Ninth Conference on Computational Natural Language
11425 Learning*, pp. 152–164. Association for Computational Linguistics.
- 11426 Carroll, L. (1865). *Alice's Adventures in Wonderland*. London: Macmillan.
- 11427 Carroll, L. (1917). *Through the looking glass: And what Alice found there*. Chicago: Rand,
11428 McNally.
- 11429 Chambers, N. and D. Jurafsky (2008). Jointly combining implicit constraints improves
11430 temporal ordering. In *Proceedings of Empirical Methods for Natural Language Processing
11431 (EMNLP)*, pp. 698–706.
- 11432 Chang, K.-W., A. Krishnamurthy, A. Agarwal, H. Daume III, and J. Langford (2015).
11433 Learning to search better than your teacher. In *Proceedings of the International Confer-
11434 ence on Machine Learning (ICML)*.
- 11435 Chang, M.-W., L. Ratinov, and D. Roth (2007). Guiding semi-supervision with constraint-
11436 driven learning. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp.
11437 280–287.
- 11438 Chang, M.-W., L.-A. Ratinov, N. Rizzolo, and D. Roth (2008). Learning and inference with
11439 constraints. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, pp.
11440 1513–1518.
- 11441 Chapman, W. W., W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan (2001). A
11442 simple algorithm for identifying negated findings and diseases in discharge summaries.
11443 *Journal of biomedical informatics* 34(5), 301–310.
- 11444 Charniak, E. (1997). Statistical techniques for natural language parsing. *AI magazine* 18(4),
11445 33–43.
- 11446 Charniak, E. and M. Johnson (2005). Coarse-to-fine n-best parsing and maxent discrimi-
11447 native reranking. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp.
11448 173–180.
- 11449 Chelba, C. and A. Acero (2006). Adaptation of maximum entropy capitalizer: Little data
11450 can help a lot. *Computer Speech & Language* 20(4), 382–399.
- 11451 Chelba, C., T. Mikolov, M. Schuster, Q. Ge, T. Brants, P. Koehn, and T. Robinson (2013).
11452 One billion word benchmark for measuring progress in statistical language modeling.
11453 *arXiv preprint arXiv:1312.3005*.
- 11454 Chen, D., J. Bolton, and C. D. Manning (2016). A thorough examination of the CNN/Daily
11455 Mail reading comprehension task. In *Proceedings of the Association for Computational
11456 Linguistics (ACL)*.

- 11457 Chen, D. and C. D. Manning (2014). A fast and accurate dependency parser using neural
11458 networks. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*,
11459 pp. 740–750.
- 11460 Chen, D. L. and R. J. Mooney (2008). Learning to sportscast: a test of grounded language
11461 acquisition. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp.
11462 128–135.
- 11463 Chen, H., S. Branavan, R. Barzilay, and D. R. Karger (2009). Content modeling using latent
11464 permutations. *Journal of Artificial Intelligence Research* 36(1), 129–163.
- 11465 Chen, M., Z. Xu, K. Weinberger, and F. Sha (2012). Marginalized denoising autoencoders
11466 for domain adaptation. In *Proceedings of the International Conference on Machine Learning
11467 (ICML)*.
- 11468 Chen, M. X., O. Firat, A. Bapna, M. Johnson, W. Macherey, G. Foster, L. Jones, N. Parmar,
11469 M. Schuster, Z. Chen, Y. Wu, and M. Hughes (2018). The best of both worlds: Combining
11470 recent advances in neural machine translation. In *Proceedings of the Association for
11471 Computational Linguistics (ACL)*.
- 11472 Chen, S. F. and J. Goodman (1999). An empirical study of smoothing techniques for lan-
11473 guage modeling. *Computer Speech & Language* 13(4), 359–393.
- 11474 Chen, T. and C. Guestrin (2016). Xgboost: A scalable tree boosting system. In *Proceedings
11475 of Knowledge Discovery and Data Mining (KDD)*, pp. 785–794.
- 11476 Chen, X., X. Qiu, C. Zhu, P. Liu, and X. Huang (2015). Long short-term memory neural
11477 networks for chinese word segmentation. In *Proceedings of Empirical Methods for Natural
11478 Language Processing (EMNLP)*, pp. 1197–1206.
- 11479 Chen, Y., S. Gilroy, A. Malletti, K. Knight, and J. May (2018). Recurrent neural networks
11480 as weighted language recognizers. In *Proceedings of the North American Chapter of the
11481 Association for Computational Linguistics (NAACL)*.
- 11482 Chen, Z. and H. Ji (2009). Graph-based event coreference resolution. In *Proceedings of
11483 the 2009 Workshop on Graph-based Methods for Natural Language Processing*, pp. 54–57.
11484 Association for Computational Linguistics.
- 11485 Cheng, X. and D. Roth (2013). Relational inference for wikification. In *Proceedings of
11486 Empirical Methods for Natural Language Processing (EMNLP)*, pp. 1787–1796.
- 11487 Chiang, D. (2007). Hierarchical phrase-based translation. *Computational Linguistics* 33(2),
11488 201–228.

- 11489 Chiang, D., J. Graehl, K. Knight, A. Pauls, and S. Ravi (2010). Bayesian inference for
11490 finite-state transducers. In *Proceedings of the North American Chapter of the Association for*
11491 *Computational Linguistics (NAACL)*, pp. 447–455.
- 11492 Chinchor, N. and P. Robinson (1997). MUC-7 named entity task definition. In *Proceedings*
11493 *of the 7th Conference on Message Understanding*, Volume 29.
- 11494 Cho, K. (2015). Natural language understanding with distributed representation.
11495 *CoRR abs/1511.07916*.
- 11496 Cho, K., B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and
11497 Y. Bengio (2014). Learning phrase representations using rnn encoder-decoder for sta-
11498 tistical machine translation. In *Proceedings of Empirical Methods for Natural Language*
11499 *Processing (EMNLP)*.
- 11500 Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton & Co.
- 11501 Chomsky, N. (1982). *Some concepts and consequences of the theory of government and binding*,
11502 Volume 6. MIT press.
- 11503 Choromanska, A., M. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun (2015). The loss
11504 surfaces of multilayer networks. In *Proceedings of Artificial Intelligence and Statistics (AIS-
11505 TATS)*, pp. 192–204.
- 11506 Christodoulopoulos, C., S. Goldwater, and M. Steedman (2010). Two decades of unsuper-
11507 vised pos induction: How far have we come? In *Proceedings of Empirical Methods for*
11508 *Natural Language Processing (EMNLP)*, pp. 575–584.
- 11509 Chu, Y.-J. and T.-H. Liu (1965). On shortest arborescence of a directed graph. *Scientia*
11510 *Sinica* 14(10), 1396–1400.
- 11511 Chung, C. and J. W. Pennebaker (2007). The psychological functions of function words.
11512 In K. Fiedler (Ed.), *Social communication*, pp. 343–359. New York and Hove: Psychology
11513 Press.
- 11514 Church, K. (2011). A pendulum swung too far. *Linguistic Issues in Language Technology* 6(5),
11515 1–27.
- 11516 Church, K. W. (2000). Empirical estimates of adaptation: the chance of two Noriega-
11517 sis closer to $p/2$ than p^2 . In *Proceedings of the International Conference on Computational*
11518 *Linguistics (COLING)*, pp. 180–186.
- 11519 Church, K. W. and P. Hanks (1990). Word association norms, mutual information, and
11520 lexicography. *Computational linguistics* 16(1), 22–29.

- 11521 Ciaramita, M. and M. Johnson (2003). Supersense tagging of unknown nouns in wordnet.
 11522 In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 168–
 11523 175.
- 11524 Clark, K. and C. D. Manning (2015). Entity-centric coreference resolution with model
 11525 stacking. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 1405–
 11526 1415.
- 11527 Clark, K. and C. D. Manning (2016). Improving coreference resolution by learning entity-
 11528 level distributed representations. In *Proceedings of the Association for Computational Lin-
 11529 guistics (ACL)*.
- 11530 Clark, P. (2015). Elementary school science and math tests as a driver for AI: take the aristo
 11531 challenge! In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, pp.
 11532 4019–4021.
- 11533 Clarke, J., D. Goldwasser, M.-W. Chang, and D. Roth (2010). Driving semantic parsing
 11534 from the world’s response. In *Proceedings of the Conference on Natural Language Learning
 11535 (CoNLL)*, pp. 18–27.
- 11536 Clarke, J. and M. Lapata (2008). Global inference for sentence compression: An integer
 11537 linear programming approach. *Journal of Artificial Intelligence Research* 31, 399–429.
- 11538 Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychologi-
 11539 cal measurement* 20(1), 37–46.
- 11540 Cohen, S. (2016). *Bayesian analysis in natural language processing*. Synthesis Lectures on
 11541 Human Language Technologies. San Rafael, CA: Morgan & Claypool Publishers.
- 11542 Cohen, S. B., K. Stratos, M. Collins, D. P. Foster, and L. Ungar (2014). Spectral learning of
 11543 latent-variable PCFGs: Algorithms and sample complexity. *Journal of Machine Learning
 11544 Research* 15, 2399–2449.
- 11545 Collier, N., C. Nobata, and J.-i. Tsujii (2000). Extracting the names of genes and gene
 11546 products with a hidden markov model. In *Proceedings of the International Conference on
 11547 Computational Linguistics (COLING)*, pp. 201–207.
- 11548 Collins, M. (1997). Three generative, lexicalised models for statistical parsing. In *Proceed-
 11549 ings of the Association for Computational Linguistics (ACL)*, pp. 16–23.
- 11550 Collins, M. (2002). Discriminative training methods for hidden markov models: theory
 11551 and experiments with perceptron algorithms. In *Proceedings of Empirical Methods for
 11552 Natural Language Processing (EMNLP)*, pp. 1–8.

- 11553 Collins, M. (2013). Notes on natural language processing. <http://www.cs.columbia.edu/~mcollins/notes-spring2013.html>.
- 11554
- 11555 Collins, M. and T. Koo (2005). Discriminative reranking for natural language parsing. *Computational Linguistics* 31(1), 25–70.
- 11556
- 11557 Collins, M. and B. Roark (2004). Incremental parsing with the perceptron algorithm. In *Proceedings of the Association for Computational Linguistics (ACL)*.
- 11558
- 11559 Collobert, R., K. Kavukcuoglu, and C. Farabet (2011). Torch7: A matlab-like environment for machine learning. Technical Report EPFL-CONF-192376, EPFL.
- 11560
- 11561 Collobert, R. and J. Weston (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 160–167.
- 11562
- 11563
- 11564 Collobert, R., J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research* 12, 2493–2537.
- 11565
- 11566
- 11567 Colton, S., J. Goodwin, and T. Veale (2012). Full-face poetry generation. In *Proceedings of the International Conference on Computational Creativity*, pp. 95–102.
- 11568
- 11569 Conneau, A., D. Kiela, H. Schwenk, L. Barrault, and A. Bordes (2017). Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 681–691.
- 11570
- 11571
- 11572 Cormen, T. H., C. E. Leiserson, R. L. Rivest, and C. Stein (2009). *Introduction to algorithms* (third ed.). MIT press.
- 11573
- 11574 Cotterell, R., H. Schütze, and J. Eisner (2016). Morphological smoothing and extrapolation of word embeddings. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 1651–1660.
- 11575
- 11576
- 11577 Coviello, L., Y. Sohn, A. D. Kramer, C. Marlow, M. Franceschetti, N. A. Christakis, and J. H. Fowler (2014). Detecting emotional contagion in massive social networks. *PloS one* 9(3), e90315.
- 11578
- 11579
- 11580 Covington, M. A. (2001). A fundamental algorithm for dependency parsing. In *Proceedings of the 39th annual ACM southeast conference*, pp. 95–102.
- 11581
- 11582 Crammer, K., O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer (2006, December). Online passive-aggressive algorithms. *The Journal of Machine Learning Research* 7, 551–585.
- 11583
- 11584

- 11585 Crammer, K. and Y. Singer (2001). Pranking with ranking. In *Neural Information Processing Systems (NIPS)*, pp. 641–647.
- 11586
- 11587 Crammer, K. and Y. Singer (2003). Ultraconservative online algorithms for multiclass problems. *The Journal of Machine Learning Research* 3, 951–991.
- 11588
- 11589 Creutz, M. and K. Lagus (2007). Unsupervised models for morpheme segmentation and 11590 morphology learning. *ACM Transactions on Speech and Language Processing (TSLP)* 4(1), 11591 3.
- 11592 Cross, J. and L. Huang (2016). Span-based constituency parsing with a structure-label 11593 system and provably optimal dynamic oracles. In *Proceedings of Empirical Methods for 11594 Natural Language Processing (EMNLP)*, pp. 1–11.
- 11595 Cucerzan, S. (2007). Large-scale named entity disambiguation based on wikipedia data. 11596 In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.
- 11597 Cui, H., R. Sun, K. Li, M.-Y. Kan, and T.-S. Chua (2005). Question answering passage 11598 retrieval using dependency relations. In *Proceedings of ACM SIGIR conference on Research 11599 and development in information retrieval*, pp. 400–407.
- 11600 Cui, Y., Z. Chen, S. Wei, S. Wang, T. Liu, and G. Hu (2017). Attention-over-attention neural 11601 networks for reading comprehension. In *Proceedings of the Association for Computational 11602 Linguistics (ACL)*.
- 11603 Culotta, A. and J. Sorensen (2004). Dependency tree kernels for relation extraction. In 11604 *Proceedings of the Association for Computational Linguistics (ACL)*.
- 11605 Culotta, A., M. Wick, and A. McCallum (2007). First-order probabilistic models for coref- 11606 erence resolution. In *Proceedings of the North American Chapter of the Association for Com- 11607 putational Linguistics (NAACL)*, pp. 81–88.
- 11608 Curry, H. B. and R. Feys (1958). *Combinatory Logic*, Volume I. Amsterdam: North Holland.
- 11609 Danescu-Niculescu-Mizil, C., M. Sudhof, D. Jurafsky, J. Leskovec, and C. Potts (2013). A 11610 computational approach to politeness with application to social factors. In *Proceedings 11611 of the Association for Computational Linguistics (ACL)*, pp. 250–259.
- 11612 Das, D., D. Chen, A. F. Martins, N. Schneider, and N. A. Smith (2014). Frame-semantic 11613 parsing. *Computational Linguistics* 40(1), 9–56.
- 11614 Daumé III, H. (2007). Frustratingly easy domain adaptation. In *Proceedings of the Associa- 11615 tion for Computational Linguistics (ACL)*.
- 11616 Daumé III, H., J. Langford, and D. Marcu (2009). Search-based structured prediction. 11617 *Machine learning* 75(3), 297–325.

- 11618 Daumé III, H. and D. Marcu (2005). A large-scale exploration of effective global features
11619 for a joint entity detection and tracking model. In *Proceedings of Empirical Methods for*
11620 *Natural Language Processing (EMNLP)*, pp. 97–104.
- 11621 Dauphin, Y. N., R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio (2014). Identifying
11622 and attacking the saddle point problem in high-dimensional non-convex optimization.
11623 In *Neural Information Processing Systems (NIPS)*, pp. 2933–2941.
- 11624 Davidson, D. (1967). The logical form of action sentences. In N. Rescher (Ed.), *The Logic of*
11625 *Decision and Action*. Pittsburgh: University of Pittsburgh Press.
- 11626 De Gispert, A. and J. B. Marino (2006). Catalan-english statistical machine translation
11627 without parallel corpus: bridging through spanish. In *Proc. of 5th International Conference*
11628 *on Language Resources and Evaluation (LREC)*, pp. 65–68. Citeseer.
- 11629 De Marneffe, M.-C. and C. D. Manning (2008). The stanford typed dependencies represen-
11630 tation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain*
11631 *Parser Evaluation*, pp. 1–8. Association for Computational Linguistics.
- 11632 Dean, J. and S. Ghemawat (2008). Mapreduce: simplified data processing on large clusters.
11633 *Communications of the ACM* 51(1), 107–113.
- 11634 Deerwester, S. C., S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman (1990).
11635 Indexing by latent semantic analysis. *Journal of the American society for information sci-
11636 ence* 41(6), 391–407.
- 11637 Dehdari, J. (2014). *A Neurophysiologically-Inspired Statistical Language Model*. Ph. D. thesis,
11638 The Ohio State University.
- 11639 Deisenroth, M. P., A. A. Faisal, and C. S. Ong (2018). *Mathematics For Machine Learning*.
11640 Cambridge University Press.
- 11641 Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incom-
11642 plete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Method-
11643 ological)*, 1–38.
- 11644 Denis, P. and J. Baldridge (2007). A ranking approach to pronoun resolution. In *Proceedings*
11645 *of the International Joint Conference on Artificial Intelligence (IJCAI)*.
- 11646 Denis, P. and J. Baldridge (2008). Specialized models and ranking for coreference resolu-
11647 tion. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp.
11648 660–669.
- 11649 Denis, P. and J. Baldridge (2009). Global joint models for coreference resolution and named
11650 entity classification. *Procesamiento del Lenguaje Natural* 42.

- 11651 Derrida, J. (1985). Des tours de babel. In J. Graham (Ed.), *Difference in translation*. Ithaca,
11652 NY: Cornell University Press.
- 11653 Dhingra, B., H. Liu, Z. Yang, W. W. Cohen, and R. Salakhutdinov (2017). Gated-attention
11654 readers for text comprehension. In *Proceedings of the Association for Computational Lin-*
11655 *guistics (ACL)*.
- 11656 Diaconis, P. and B. Skyrms (2017). *Ten Great Ideas About Chance*. Princeton University
11657 Press.
- 11658 Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classifica-
11659 tion learning algorithms. *Neural computation* 10(7), 1895–1923.
- 11660 Dietterich, T. G., R. H. Lathrop, and T. Lozano-Pérez (1997). Solving the multiple instance
11661 problem with axis-parallel rectangles. *Artificial intelligence* 89(1), 31–71.
- 11662 Dimitrova, L., N. Ide, V. Petkevic, T. Erjavec, H. J. Kaalep, and D. Tufis (1998). Multext-
11663 east: Parallel and comparable corpora and lexicons for six central and eastern euro-
11664 pean languages. In *Proceedings of the International Conference on Computational Linguistics*
11665 (*COLING*), pp. 315–319.
- 11666 Doddington, G. R., A. Mitchell, M. A. Przybocki, L. A. Ramshaw, S. Strassel, and R. M.
11667 Weischedel (2004). The automatic content extraction (ACE) program-tasks, data, and
11668 evaluation. In *Proceedings of the Language Resources and Evaluation Conference*, pp. 837–
11669 840.
- 11670 dos Santos, C., B. Xiang, and B. Zhou (2015). Classifying relations by ranking with con-
11671 volutional neural networks. In *Proceedings of the Association for Computational Linguistics*
11672 (*ACL*), pp. 626–634.
- 11673 Dowty, D. (1991). Thematic proto-roles and argument selection. *Language*, 547–619.
- 11674 Dredze, M., P. McNamee, D. Rao, A. Gerber, and T. Finin (2010). Entity disambiguation
11675 for knowledge base population. In *Proceedings of the International Conference on Compu-*
11676 *tational Linguistics (COLING)*, pp. 277–285.
- 11677 Dredze, M., M. J. Paul, S. Bergsma, and H. Tran (2013). Carmen: A Twitter geolocation
11678 system with applications to public health. In *AAAI workshop on expanding the boundaries*
11679 *of health informatics using AI (HIAI)*, pp. 20–24.
- 11680 Dreyfus, H. L. (1992). *What computers still can't do: a critique of artificial reason*. MIT press.
- 11681 Du, L., W. Buntine, and M. Johnson (2013). Topic segmentation with a structured topic
11682 model. In *Proceedings of the North American Chapter of the Association for Computational*
11683 *Linguistics (NAACL)*, pp. 190–200.

- 11684 Duchi, J., E. Hazan, and Y. Singer (2011). Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research* 12, 2121–2159.
- 11685
- 11686 Dunietz, J., L. Levin, and J. Carbonell (2017). The because corpus 2.0: Annotating causality and overlapping relations. In *Proceedings of the Linguistic Annotation Workshop*.
- 11687
- 11688 Durrett, G., T. Berg-Kirkpatrick, and D. Klein (2016). Learning-based single-document summarization with compression and anaphoricity constraints. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 1998–2008.
- 11689
- 11690
- 11691 Durrett, G. and D. Klein (2013). Easy victories and uphill battles in coreference resolution. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.
- 11692
- 11693 Durrett, G. and D. Klein (2015). Neural CRF parsing. In *Proceedings of the Association for Computational Linguistics (ACL)*.
- 11694
- 11695 Dyer, C., M. Ballesteros, W. Ling, A. Matthews, and N. A. Smith (2015). Transition-based dependency parsing with stack long short-term memory. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 334–343.
- 11696
- 11697
- 11698 Dyer, C., A. Kuncoro, M. Ballesteros, and N. A. Smith (2016). Recurrent neural network grammars. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 199–209.
- 11699
- 11700
- 11701 Edmonds, J. (1967). Optimum branchings. *Journal of Research of the National Bureau of Standards B* 71(4), 233–240.
- 11702
- 11703 Efron, B. and R. J. Tibshirani (1993). An introduction to the bootstrap: Monographs on statistics and applied probability, vol. 57. New York and London: Chapman and Hall/CRC.
- 11704
- 11705 Eisenstein, J. (2009). Hierarchical text segmentation from multi-scale lexical cohesion. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.
- 11706
- 11707
- 11708 Eisenstein, J. and R. Barzilay (2008). Bayesian unsupervised topic segmentation. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.
- 11709
- 11710 Eisner, J. (1997). State-of-the-art algorithms for minimum spanning trees: A tutorial discussion. <https://www.cs.jhu.edu/~jason/papers/eisner.mst-tutorial.pdf>.
- 11711
- 11712
- 11713 Eisner, J. (2000). Bilexical grammars and their cubic-time parsing algorithms. In *Advances in probabilistic and other parsing technologies*, pp. 29–61. Springer.
- 11714
- 11715 Eisner, J. (2002). Parameter estimation for probabilistic finite-state transducers. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 1–8.
- 11716

- 11717 Eisner, J. (2016). Inside-outside and forward-backward algorithms are just backprop. In
11718 *Proceedings of the Workshop on Structured Prediction for NLP*, pp. 1–17.
- 11719 Eisner, J. M. (1996). Three new probabilistic models for dependency parsing: An explo-
11720 ration. In *Proceedings of the International Conference on Computational Linguistics (COL-*
11721 *ING)*, pp. 340–345.
- 11722 Ekman, P. (1992). Are there basic emotions? *Psychological Review* 99(3), 550–553.
- 11723 Elman, J. L. (1990). Finding structure in time. *Cognitive science* 14(2), 179–211.
- 11724 Elman, J. L., E. A. Bates, M. H. Johnson, A. Karmiloff-Smith, D. Parisi, and K. Plunkett
11725 (1998). *Rethinking innateness: A connectionist perspective on development*, Volume 10. MIT
11726 press.
- 11727 Elsner, M. and E. Charniak (2010). Disentangling chat. *Computational Linguistics* 36(3),
11728 389–409.
- 11729 Esuli, A. and F. Sebastiani (2006). SentiWordNet: A publicly available lexical resource for
11730 opinion mining. In *Proceedings of the Language Resources and Evaluation Conference*, pp.
11731 417–422.
- 11732 Etzioni, O., A. Fader, J. Christensen, S. Soderland, and M. Mausam (2011). Open informa-
11733 tion extraction: The second generation. In *Proceedings of the International Joint Conference
11734 on Artificial Intelligence (IJCAI)*, pp. 3–10.
- 11735 Faruqui, M., J. Dodge, S. K. Jauhar, C. Dyer, E. Hovy, and N. A. Smith (2015). Retrofitting
11736 word vectors to semantic lexicons. In *Proceedings of the North American Chapter of the
11737 Association for Computational Linguistics (NAACL)*.
- 11738 Faruqui, M. and C. Dyer (2014). Improving vector space word representations using mul-
11739 tilingual correlation. In *Proceedings of the European Chapter of the Association for Compu-
11740 tational Linguistics (EACL)*, pp. 462–471.
- 11741 Faruqui, M., R. McDonald, and R. Soricut (2016). Morpho-syntactic lexicon generation
11742 using graph-based semi-supervised learning. *Transactions of the Association for Compu-
11743 tational Linguistics* 4, 1–16.
- 11744 Faruqui, M., Y. Tsvetkov, P. Rastogi, and C. Dyer (2016). Problems with evaluation of word
11745 embeddings using word similarity tasks. In *Proceedings of the 1st Workshop on Evaluating
11746 Vector-Space Representations for NLP*, pp. 30–35. Association for Computational Linguis-
11747 tics.
- 11748 Fellbaum, C. (2010). *WordNet*. Springer.

- 11749 Feng, V. W., Z. Lin, and G. Hirst (2014). The impact of deep hierarchical discourse structures in the evaluation of text coherence. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pp. 940–949.
- 11752 Feng, X., L. Huang, D. Tang, H. Ji, B. Qin, and T. Liu (2016). A language-independent neural network for event detection. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 66–71.
- 11755 Fernandes, E. R., C. N. dos Santos, and R. L. Milidiú (2014). Latent trees for coreference resolution. *Computational Linguistics*.
- 11757 Ferrucci, D., E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. A. Kalyanpur, A. Lally, J. W. Murdock, E. Nyberg, J. Prager, et al. (2010). Building Watson: An overview of the DeepQA project. *AI magazine* 31(3), 59–79.
- 11760 Ficler, J. and Y. Goldberg (2017). Controlling linguistic style aspects in neural language generation. In *Proceedings of the Workshop on Stylistic Variation*, pp. 94–104. Association for Computational Linguistics.
- 11763 Filippova, K. and M. Strube (2008). Sentence fusion via dependency graph compression. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 177–185.
- 11766 Fillmore, C. J. (1968). The case for case. In E. Bach and R. Harms (Eds.), *Universals in linguistic theory*. Holt, Rinehart, and Winston.
- 11768 Fillmore, C. J. (1976). Frame semantics and the nature of language. *Annals of the New York Academy of Sciences* 280(1), 20–32.
- 11770 Fillmore, C. J. and C. Baker (2009). A frames approach to semantic analysis. In *The Oxford Handbook of Linguistic Analysis*. Oxford University Press.
- 11772 Finkel, J. R., T. Grenager, and C. Manning (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 363–370.
- 11775 Finkel, J. R., T. Grenager, and C. D. Manning (2007). The infinite tree. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 272–279.
- 11777 Finkel, J. R., A. Kleeman, and C. D. Manning (2008). Efficient, feature-based, conditional random field parsing. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 959–967.
- 11780 Finkel, J. R. and C. Manning (2009). Hierarchical bayesian domain adaptation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 602–610.

- 11783 Finkel, J. R. and C. D. Manning (2008). Enforcing transitivity in coreference resolution. In
11784 *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 45–48.
- 11785 Finkelstein, L., E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin
11786 (2002). Placing search in context: The concept revisited. *ACM Transactions on Information
11787 Systems* 20(1), 116–131.
- 11788 Firth, J. R. (1957). *Papers in Linguistics 1934-1951*. Oxford University Press.
- 11789 Flanigan, J., S. Thomson, J. Carbonell, C. Dyer, and N. A. Smith (2014). A discriminative
11790 graph-based parser for the abstract meaning representation. In *Proceedings of the
11791 Association for Computational Linguistics (ACL)*, pp. 1426–1436.
- 11792 Foltz, P. W., W. Kintsch, and T. K. Landauer (1998). The measurement of textual coherence
11793 with latent semantic analysis. *Discourse processes* 25(2-3), 285–307.
- 11794 Fordyce, C. (2007). Overview of the IWSLT 2007 evaluation campaign. In *International
11795 Workshop on Spoken Language Translation (IWSLT) 2007*.
- 11796 Forsyth, E. N. and C. H. Martell (2007). Lexical and discourse analysis of online chat
11797 dialog. In *International Conference on Semantic Computing*, pp. 19–26. IEEE.
- 11798 Fort, K., G. Adda, and K. B. Cohen (2011). Amazon mechanical turk: Gold mine or coal
11799 mine? *Computational Linguistics* 37(2), 413–420.
- 11800 Fox, H. (2002). Phrasal cohesion and statistical machine translation. In *Proceedings of
11801 Empirical Methods for Natural Language Processing (EMNLP)*, pp. 304–3111.
- 11802 Francis, W. and H. Kucera (1982). *Frequency analysis of English usage*. Houghton Mifflin
11803 Company.
- 11804 Francis, W. N. (1964). A standard sample of present-day English for use with digital
11805 computers. Report to the U.S Office of Education on Cooperative Research Project No.
11806 E-007.
- 11807 Freund, Y. and R. E. Schapire (1999). Large margin classification using the perceptron
11808 algorithm. *Machine learning* 37(3), 277–296.
- 11809 Fromkin, V., R. Rodman, and N. Hyams (2013). *An introduction to language*. Cengage
11810 Learning.
- 11811 Fundel, K., R. Küffner, and R. Zimmer (2007). Relex – relation extraction using depen-
11812 dency parse trees. *Bioinformatics* 23(3), 365–371.

- 11813 Gabow, H. N., Z. Galil, T. Spencer, and R. E. Tarjan (1986). Efficient algorithms for finding
11814 minimum spanning trees in undirected and directed graphs. *Combinatorica* 6(2), 109–
11815 122.
- 11816 Gabrilovich, E. and S. Markovitch (2007). Computing semantic relatedness using
11817 wikipedia-based explicit semantic analysis. In *Proceedings of the International Joint Con-*
11818 *ference on Artificial Intelligence (IJCAI)*, Volume 7, pp. 1606–1611.
- 11819 Gage, P. (1994). A new algorithm for data compression. *The C Users Journal* 12(2), 23–38.
- 11820 Gale, W. A., K. W. Church, and D. Yarowsky (1992). One sense per discourse. In *Pro-*
11821 *ceedings of the workshop on Speech and Natural Language*, pp. 233–237. Association for
11822 Computational Linguistics.
- 11823 Galley, M., M. Hopkins, K. Knight, and D. Marcu (2004). What's in a translation rule? In
11824 *Proceedings of the North American Chapter of the Association for Computational Linguistics*
11825 (NAACL), pp. 273–280.
- 11826 Galley, M., K. R. McKeown, E. Fosler-Lussier, and H. Jing (2003). Discourse segmentation
11827 of multi-party conversation. In *Proceedings of the Association for Computational Linguistics*
11828 (ACL).
- 11829 Ganchev, K. and M. Dredze (2008). Small statistical models by random feature mixing. In
11830 *Proceedings of Workshop on Mobile Language Processing*, pp. 19–20. Association of Compu-
11831 tational Linguistics.
- 11832 Ganchev, K., J. Graça, J. Gillenwater, and B. Taskar (2010). Posterior regularization for
11833 structured latent variable models. *The Journal of Machine Learning Research* 11, 2001–
11834 2049.
- 11835 Ganin, Y., E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand,
11836 and V. Lempitsky (2016). Domain-adversarial training of neural networks. *The Journal*
11837 *of Machine Learning Research* 17(59), 1–35.
- 11838 Gao, J., G. Andrew, M. Johnson, and K. Toutanova (2007). A comparative study of param-
11839 eter estimation methods for statistical natural language processing. In *Proceedings of the*
11840 *Association for Computational Linguistics (ACL)*, pp. 824–831.
- 11841 Garg, N., L. Schiebinger, D. Jurafsky, and J. Zou (2018). Word embeddings quantify
11842 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sci-*
11843 *ences* 115(16), E3635–E3644.
- 11844 Gatt, A. and E. Krahmer (2018). Survey of the state of the art in natural language genera-
11845 tion: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research* 61,
11846 65–170.

- 11847 Ge, D., X. Jiang, and Y. Ye (2011). A note on the complexity of l_p minimization. *Mathematical programming* 129(2), 285–299.
- 11848
- 11849 Ge, N., J. Hale, and E. Charniak (1998). A statistical approach to anaphora resolution. In *Proceedings of the sixth workshop on very large corpora*, Volume 71, pp. 76.
- 11850
- 11851 Ge, R., F. Huang, C. Jin, and Y. Yuan (2015). Escaping from saddle points — online stochastic gradient for tensor decomposition. In P. Grünwald, E. Hazan, and S. Kale (Eds.), *Proceedings of the Conference On Learning Theory (COLT)*.
- 11852
- 11853
- 11854 Ge, R. and R. J. Mooney (2005). A statistical semantic parser that integrates syntax and semantics. In *Proceedings of the Conference on Natural Language Learning (CoNLL)*, pp. 9–16.
- 11855
- 11856
- 11857 Geach, P. T. (1962). *Reference and generality: An examination of some medieval and modern theories*. Cornell University Press.
- 11858
- 11859 Gehring, J., M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin (2017). Convolutional sequence to sequence learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 1243–1252.
- 11860
- 11861
- 11862 Gildea, D. and D. Jurafsky (2002). Automatic labeling of semantic roles. *Computational linguistics* 28(3), 245–288.
- 11863
- 11864 Gimpel, K., N. Schneider, B. O’Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogoatama, J. Flanigan, and N. A. Smith (2011). Part-of-speech tagging for Twitter: annotation, features, and experiments. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 42–47.
- 11865
- 11866
- 11867
- 11868 Glass, J., T. J. Hazen, S. Cyphers, I. Malioutov, D. Huynh, and R. Barzilay (2007). Recent progress in the mit spoken lecture processing project. In *Eighth Annual Conference of the International Speech Communication Association*.
- 11869
- 11870
- 11871 Glorot, X. and Y. Bengio (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of Artificial Intelligence and Statistics (AISTATS)*, pp. 249–256.
- 11872
- 11873
- 11874 Glorot, X., A. Bordes, and Y. Bengio (2011). Deep sparse rectifier networks. In *Proceedings of Artificial Intelligence and Statistics (AISTATS)*, pp. 315–323.
- 11875
- 11876 Godfrey, J. J., E. C. Holliman, and J. McDaniel (1992). Switchboard: Telephone speech corpus for research and development. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 517–520. IEEE.
- 11877
- 11878

- 11879 Goldberg, Y. (2017a, June). An adversarial review of “adversarial generation of
11880 natural language”. <https://medium.com/@yoav.goldberg/an-adversarial-review-of->
11881 adversarial-generation-of-natural-language-409ac3378bd7.
- 11882 Goldberg, Y. (2017b). *Neural Network Methods for Natural Language Processing*. Synthesis
11883 Lectures on Human Language Technologies. Morgan & Claypool Publishers.
- 11884 Goldberg, Y. and M. Elhadad (2010). An efficient algorithm for easy-first non-directional
11885 dependency parsing. In *Proceedings of the North American Chapter of the Association for*
11886 *Computational Linguistics* (NAACL), pp. 742–750.
- 11887 Goldberg, Y. and J. Nivre (2012). A dynamic oracle for arc-eager dependency parsing.
11888 In *Proceedings of the International Conference on Computational Linguistics* (COLING), pp.
11889 959–976.
- 11890 Goldberg, Y., K. Zhao, and L. Huang (2013). Efficient implementation of beam-search
11891 incremental parsers. In *Proceedings of the Association for Computational Linguistics* (ACL),
11892 pp. 628–633.
- 11893 Goldwater, S. and T. Griffiths (2007). A fully bayesian approach to unsupervised part-of-
11894 speech tagging. In *Proceedings of the Association for Computational Linguistics* (ACL).
- 11895 Gonçalo Oliveira, H. R., F. A. Cardoso, and F. C. Pereira (2007). Tra-la-lyrics: An approach
11896 to generate text based on rhythm. In *Proceedings of the 4th. International Joint Workshop*
11897 *on Computational Creativity*. A. Cardoso and G. Wiggins.
- 11898 Goodfellow, I., Y. Bengio, and A. Courville (2016). *Deep learning*. MIT Press.
- 11899 Goodman, J. T. (2001). A bit of progress in language modeling. *Computer Speech & Lan-*
11900 *guage* 15(4), 403–434.
- 11901 Gouws, S., D. Metzler, C. Cai, and E. Hovy (2011). Contextual bearing on linguistic varia-
11902 tion in social media. In *Proceedings of the Workshop on Language and Social Media*. Associa-
11903 tion for Computational Linguistics.
- 11904 Goyal, A., H. Daumé III, and S. Venkatasubramanian (2009). Streaming for large scale
11905 NLP: Language modeling. In *Proceedings of the North American Chapter of the Association*
11906 *for Computational Linguistics* (NAACL), pp. 512–520.
- 11907 Graves, A. (2012). Sequence transduction with recurrent neural networks. In *Proceedings*
11908 *of the International Conference on Machine Learning* (ICML).
- 11909 Graves, A. and N. Jaitly (2014). Towards end-to-end speech recognition with recur-
11910 rent neural networks. In *Proceedings of the International Conference on Machine Learning*
11911 *(ICML)*, pp. 1764–1772.

- 11912 Graves, A. and J. Schmidhuber (2005). Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks* 18(5), 602–610.
- 11913
- 11914 Grice, H. P. (1975). Logic and conversation. In P. Cole and J. L. Morgan (Eds.), *Syntax and Semantics Volume 3: Speech Acts*, pp. 41–58. Academic Press.
- 11915
- 11916 Grishman, R. (2012). Information extraction: Capabilities and challenges. Notes prepared for the 2012 International Winter School in Language and Speech Technologies, Rovira i Virgili University, Tarragona, Spain.
- 11917
- 11918
- 11919 Grishman, R. (2015). Information extraction. *IEEE Intelligent Systems* 30(5), 8–15.
- 11920 Grishman, R., C. Macleod, and J. Sterling (1992). Evaluating parsing strategies using standardized parse files. In *Proceedings of the third conference on Applied natural language processing*, pp. 156–161. Association for Computational Linguistics.
- 11921
- 11922
- 11923 Grishman, R. and B. Sundheim (1996). Message understanding conference-6: A brief history. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pp. 466–471.
- 11924
- 11925
- 11926 Groenendijk, J. and M. Stokhof (1991). Dynamic predicate logic. *Linguistics and philosophy* 14(1), 39–100.
- 11927
- 11928 Grosz, B. J. (1979). Focusing and description in natural language dialogues. Technical report, SRI International.
- 11929
- 11930 Grosz, B. J., S. Weinstein, and A. K. Joshi (1995). Centering: A framework for modeling the local coherence of discourse. *Computational linguistics* 21(2), 203–225.
- 11931
- 11932 Gu, J., Z. Lu, H. Li, and V. O. Li (2016). Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 1631–1640.
- 11933
- 11934
- 11935 Gulcehre, C., S. Ahn, R. Nallapati, B. Zhou, and Y. Bengio (2016). Pointing the unknown words. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 140–149.
- 11936
- 11937 Gutmann, M. U. and A. Hyvärinen (2012). Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *The Journal of Machine Learning Research* 13(1), 307–361.
- 11938
- 11939
- 11940 Haghghi, A. and D. Klein (2007). Unsupervised coreference resolution in a nonparametric bayesian model. In *Proceedings of the Association for Computational Linguistics (ACL)*.
- 11941
- 11942 Haghghi, A. and D. Klein (2009). Simple coreference resolution with rich syntactic and semantic features. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 1152–1161.
- 11943
- 11944

- 11945 Haghghi, A. and D. Klein (2010). Coreference resolution in a modular, entity-centered
11946 model. In *Proceedings of the North American Chapter of the Association for Computational*
11947 *Linguistics (NAACL)*, pp. 385–393.
- 11948 Hajič, J. and B. Hladká (1998). Tagging inflective languages: Prediction of morphological
11949 categories for a rich, structured tagset. In *Proceedings of the Association for Computational*
11950 *Linguistics (ACL)*, pp. 483–490.
- 11951 Halliday, M. and R. Hasan (1976). *Cohesion in English*. London: Longman.
- 11952 Hammerton, J. (2003). Named entity recognition with long short-term memory. In *Pro-
11953 ceedings of the Conference on Natural Language Learning (CoNLL)*, pp. 172–175.
- 11954 Han, X. and L. Sun (2012). An entity-topic model for entity linking. In *Proceedings of
11955 Empirical Methods for Natural Language Processing (EMNLP)*, pp. 105–115.
- 11956 Han, X., L. Sun, and J. Zhao (2011). Collective entity linking in web text: a graph-based
11957 method. In *Proceedings of ACM SIGIR conference on Research and development in informa-
11958 tion retrieval*, pp. 765–774.
- 11959 Hannak, A., E. Anderson, L. F. Barrett, S. Lehmann, A. Mislove, and M. Riedewald (2012).
11960 Tweetin’ in the rain: Exploring societal-scale effects of weather on mood. In *Proceedings
11961 of the International Conference on Web and Social Media (ICWSM)*.
- 11962 Hardmeier, C. (2012). Discourse in statistical machine translation. a survey and a case
11963 study. *Discours* (11).
- 11964 Haspelmath, M. and A. Sims (2013). *Understanding morphology*. Routledge.
- 11965 Hastie, T., R. Tibshirani, and J. Friedman (2009). *The elements of statistical learning* (Second
11966 ed.). New York: Springer.
- 11967 Hatzivassiloglou, V. and K. R. McKeown (1997). Predicting the semantic orientation of
11968 adjectives. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 174–
11969 181.
- 11970 Hayes, A. F. and K. Krippendorff (2007). Answering the call for a standard reliability
11971 measure for coding data. *Communication methods and measures* 1(1), 77–89.
- 11972 He, H., A. Balakrishnan, M. Eric, and P. Liang (2017). Learning symmetric collaborative
11973 dialogue agents with dynamic knowledge graph embeddings. In *Proceedings of the As-
11974 sociation for Computational Linguistics (ACL)*, pp. 1766–1776.
- 11975 He, K., X. Zhang, S. Ren, and J. Sun (2015). Delving deep into rectifiers: Surpassing
11976 human-level performance on imagenet classification. In *Proceedings of the International
11977 Conference on Computer Vision (ICCV)*, pp. 1026–1034.

- 11978 He, K., X. Zhang, S. Ren, and J. Sun (2016). Deep residual learning for image recognition.
 11979 In *Proceedings of the International Conference on Computer Vision (ICCV)*, pp. 770–778.
- 11980 He, L., K. Lee, M. Lewis, and L. Zettlemoyer (2017). Deep semantic role labeling: What
 11981 works and what’s next. In *Proceedings of the Association for Computational Linguistics*
 11982 (*ACL*).
- 11983 He, Z., S. Liu, M. Li, M. Zhou, L. Zhang, and H. Wang (2013). Learning entity repre-
 11984 sentation for entity disambiguation. In *Proceedings of the Association for Computational*
 11985 *Linguistics (ACL)*, pp. 30–34.
- 11986 Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In
 11987 *Proceedings of the International Conference on Computational Linguistics (COLING)*, pp. 539–
 11988 545.
- 11989 Hearst, M. A. (1997). Texttiling: Segmenting text into multi-paragraph subtopic passages.
 11990 *Computational linguistics* 23(1), 33–64.
- 11991 Heerschap, B., F. Goossen, A. Hogenboom, F. Frasincar, U. Kaymak, and F. de Jong (2011).
 11992 Polarity analysis of texts using discourse structure. In *Proceedings of the 20th ACM inter-*
 11993 *national conference on Information and knowledge management*, pp. 1061–1070. ACM.
- 11994 Henderson, J. (2004). Discriminative training of a neural network statistical parser. In
 11995 *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 95–102.
- 11996 Hendrickx, I., S. N. Kim, Z. Kozareva, P. Nakov, D. Ó Séaghdha, S. Padó, M. Pennacchiotti,
 11997 L. Romano, and S. Szpakowicz (2009). SemEval-2010 task 8: Multi-way classification of
 11998 semantic relations between pairs of nominals. In *Proceedings of the Workshop on Semantic*
 11999 *Evaluations: Recent Achievements and Future Directions*, pp. 94–99. Association for Com-
 12000 *computational Linguistics*.
- 12001 Hermann, K. M., T. Kočiský, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and
 12002 P. Blunsom (2015). Teaching machines to read and comprehend. In *Advances in Neu-*
 12003 *ral Information Processing Systems*, pp. 1693–1701.
- 12004 Hernault, H., H. Prendinger, D. A. duVerle, and M. Ishizuka (2010). HILDA: A discourse
 12005 parser using support vector machine classification. *Dialogue and Discourse* 1(3), 1–33.
- 12006 Hill, F., A. Bordes, S. Chopra, and J. Weston (2016). The goldilocks principle: Reading
 12007 children’s books with explicit memory representations. In *Proceedings of the International*
 12008 *Conference on Learning Representations (ICLR)*.
- 12009 Hill, F., K. Cho, and A. Korhonen (2016). Learning distributed representations of sentences
 12010 from unlabelled data. In *Proceedings of the North American Chapter of the Association for*
 12011 *Computational Linguistics (NAACL)*.

- 12012 Hill, F., R. Reichart, and A. Korhonen (2015). Simlex-999: Evaluating semantic models
12013 with (genuine) similarity estimation. *Computational Linguistics* 41(4), 665–695.
- 12014 Hindle, D. and M. Rooth (1993). Structural ambiguity and lexical relations. *Computational
12015 linguistics* 19(1), 103–120.
- 12016 Hirao, T., Y. Yoshida, M. Nishino, N. Yasuda, and M. Nagata (2013). Single-document
12017 summarization as a tree knapsack problem. In *Proceedings of Empirical Methods for Nat-
12018 ural Language Processing (EMNLP)*, pp. 1515–1520.
- 12019 Hirschman, L. and R. Gaizauskas (2001). Natural language question answering: the view
12020 from here. *natural language engineering* 7(4), 275–300.
- 12021 Hirschman, L., M. Light, E. Breck, and J. D. Burger (1999). Deep read: A reading compre-
12022 hension system. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp.
12023 325–332.
- 12024 Hobbs, J. R. (1978). Resolving pronoun references. *Lingua* 44(4), 311–338.
- 12025 Hobbs, J. R., D. Appelt, J. Bear, D. Israel, M. Kameyama, M. Stickel, and M. Tyson (1997).
12026 Fastus: A cascaded finite-state transducer for extracting information from natural-
12027 language text. *Finite-state language processing*, 383–406.
- 12028 Hochreiter, S. and J. Schmidhuber (1997). Long short-term memory. *Neural computa-
12029 tion* 9(8), 1735–1780.
- 12030 Hockenmaier, J. and M. Steedman (2007). CCGbank: a corpus of CCG derivations and de-
12031 pendency structures extracted from the Penn Treebank. *Computational Linguistics* 33(3),
12032 355–396.
- 12033 Hoffart, J., M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva,
12034 S. Thater, and G. Weikum (2011). Robust disambiguation of named entities in text. In
12035 *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 782–792.
- 12036 Hoffmann, R., C. Zhang, X. Ling, L. Zettlemoyer, and D. S. Weld (2011). Knowledge-based
12037 weak supervision for information extraction of overlapping relations. In *Proceedings of
12038 the Association for Computational Linguistics (ACL)*, pp. 541–550.
- 12039 Holmstrom, L. and P. Koistinen (1992). Using additive noise in back-propagation training.
12040 *IEEE Transactions on Neural Networks* 3(1), 24–38.
- 12041 Hovy, D., S. Spruit, M. Mitchell, E. M. Bender, M. Strube, and H. Wallach (2017). Proceed-
12042 ings of the first acl workshop on ethics in natural language processing. In *Proceedings of
12043 the First ACL Workshop on Ethics in Natural Language Processing*. Association for Compu-
12044 tational Linguistics.

- 12045 Hovy, E. and J. Lavid (2010). Towards a 'science' of corpus annotation: a new method-
12046 ological challenge for corpus linguistics. *International journal of translation* 22(1), 13–36.
- 12047 Hsu, D., S. M. Kakade, and T. Zhang (2012). A spectral algorithm for learning hidden
12048 markov models. *Journal of Computer and System Sciences* 78(5), 1460–1480.
- 12049 Hu, M. and B. Liu (2004). Mining and summarizing customer reviews. In *Proceedings of*
12050 *Knowledge Discovery and Data Mining (KDD)*, pp. 168–177.
- 12051 Hu, Z., Z. Yang, X. Liang, R. Salakhutdinov, and E. P. Xing (2017). Toward controlled gen-
12052 eration of text. In *Proceedings of the International Conference on Machine Learning (ICML)*,
12053 pp. 1587–1596.
- 12054 Huang, F. and A. Yates (2012). Biased representation learning for domain adaptation. In
12055 *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 1313–1323.
- 12056 Huang, L. and D. Chiang (2007). Forest rescoring: Faster decoding with integrated lan-
12057 guage models. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp.
12058 144–151.
- 12059 Huang, L., S. Fayong, and Y. Guo (2012). Structured perceptron with inexact search. In
12060 *Proceedings of the North American Chapter of the Association for Computational Linguistics*
12061 (NAACL), pp. 142–151.
- 12062 Huang, Y. (2015). *Pragmatics* (Second ed.). Oxford Textbooks in Linguistics. Oxford Uni-
12063 versity Press.
- 12064 Huang, Z., W. Xu, and K. Yu (2015). Bidirectional LSTM-CRF models for sequence tagging.
12065 *arXiv preprint arXiv:1508.01991*.
- 12066 Huddleston, R. and G. K. Pullum (2005). *A student's introduction to English grammar*. Cam-
12067 bridge University Press.
- 12068 Huffman, D. A. (1952). A method for the construction of minimum-redundancy codes.
12069 *Proceedings of the IRE* 40(9), 1098–1101.
- 12070 Humphreys, K., R. Gaizauskas, and S. Azzam (1997). Event coreference for information
12071 extraction. In *Proceedings of a Workshop on Operational Factors in Practical, Robust Anaphora*
12072 *Resolution for Unrestricted Texts*, pp. 75–81. Association for Computational Linguistics.
- 12073 Ide, N. and Y. Wilks (2006). Making sense about sense. In *Word sense disambiguation*, pp.
12074 47–73. Springer.
- 12075 Ioffe, S. and C. Szegedy (2015). Batch normalization: Accelerating deep network train-
12076 ing by reducing internal covariate shift. In *Proceedings of the International Conference on*
12077 *Machine Learning (ICML)*, pp. 448–456.

- 12078 Isozaki, H., T. Hirao, K. Duh, K. Sudoh, and H. Tsukada (2010). Automatic evaluation
12079 of translation quality for distant language pairs. In *Proceedings of Empirical Methods for*
12080 *Natural Language Processing (EMNLP)*, pp. 944–952.
- 12081 Ivanova, A., S. Oepen, L. Øvrelid, and D. Flickinger (2012). Who did what to whom? a
12082 contrastive study of syntacto-semantic dependencies. In *Proceedings of the Sixth Linguis-*
12083 *tic Annotation Workshop*, pp. 2–11. Association for Computational Linguistics.
- 12084 Iyyer, M., V. Manjunatha, J. Boyd-Graber, and H. Daumé III (2015). Deep unordered com-
12085 position rivals syntactic methods for text classification. In *Proceedings of the Association*
12086 *for Computational Linguistics (ACL)*, pp. 1681–1691.
- 12087 James, G., D. Witten, T. Hastie, and R. Tibshirani (2013). *An introduction to statistical learn-*
12088 *ing*, Volume 112. Springer.
- 12089 Janin, A., D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau,
12090 E. Shriberg, A. Stolcke, et al. (2003). The ICSI meeting corpus. In *Proceedings of the*
12091 *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.
- 12092 Jean, S., K. Cho, R. Memisevic, and Y. Bengio (2015). On using very large target vocab-
12093 uary for neural machine translation. In *Proceedings of the Association for Computational*
12094 *Linguistics (ACL)*, pp. 1–10.
- 12095 Jeong, M., C.-Y. Lin, and G. G. Lee (2009). Semi-supervised speech act recognition in
12096 emails and forums. In *Proceedings of Empirical Methods for Natural Language Processing*
12097 (*EMNLP*), pp. 1250–1259.
- 12098 Ji, H. and R. Grishman (2011). Knowledge base population: Successful approaches and
12099 challenges. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 1148–
12100 1158.
- 12101 Ji, Y., T. Cohn, L. Kong, C. Dyer, and J. Eisenstein (2015). Document context language
12102 models. In *International Conference on Learning Representations, Workshop Track*, Volume
12103 abs/1511.03962.
- 12104 Ji, Y. and J. Eisenstein (2014). Representation learning for text-level discourse parsing. In
12105 *Proceedings of the Association for Computational Linguistics (ACL)*.
- 12106 Ji, Y. and J. Eisenstein (2015). One vector is not enough: Entity-augmented distributional
12107 semantics for discourse relations. *Transactions of the Association for Computational Lin-*
12108 *guistics (TACL)*.
- 12109 Ji, Y., G. Haffari, and J. Eisenstein (2016). A latent variable recurrent neural network for
12110 discourse relation language models. In *Proceedings of the North American Chapter of the*
12111 *Association for Computational Linguistics (NAACL)*.

- 12112 Ji, Y. and N. A. Smith (2017). Neural discourse structure for text categorization. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 996–1005.
- 12113
- 12114 Ji, Y., C. Tan, S. Martschat, Y. Choi, and N. A. Smith (2017). Dynamic entity representations
12115 in neural language models. In *Proceedings of Empirical Methods for Natural Language
12116 Processing (EMNLP)*, pp. 1831–1840.
- 12117 Jiang, L., M. Yu, M. Zhou, X. Liu, and T. Zhao (2011). Target-dependent twitter sentiment
12118 classification. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp.
12119 151–160.
- 12120 Jing, H. (2000). Sentence reduction for automatic text summarization. In *Proceedings of
12121 the sixth conference on Applied natural language processing*, pp. 310–315. Association for
12122 Computational Linguistics.
- 12123 Joachims, T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of
12124 Knowledge Discovery and Data Mining (KDD)*, pp. 133–142.
- 12125 Jockers, M. L. (2015). Revealing sentiment and plot arcs with the syuzhet package. <http://www.matthewjockers.net/2015/02/02/syuzhet/>.
- 12126
- 12127 Johnson, A. E., T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody,
12128 P. Szolovits, L. A. Celi, and R. G. Mark (2016). MIMIC-III, a freely accessible critical care
12129 database. *Scientific data* 3, 160035.
- 12130 Johnson, M. (1998). PCFG models of linguistic tree representations. *Computational Lin-
12131 guistics* 24(4), 613–632.
- 12132 Johnson, R. and T. Zhang (2017). Deep pyramid convolutional neural networks for text
12133 categorization. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp.
12134 562–570.
- 12135 Joshi, A. K. (1985). Tree adjoining grammars: How much context-sensitivity is required
12136 to provide reasonable structural descriptions? In *Natural Language Processing – Theoret-
12137 ical, Computational and Psychological Perspectives*. New York, NY: Cambridge University
12138 Press.
- 12139 Joshi, A. K. and Y. Schabes (1997). Tree-adjoining grammars. In *Handbook of formal lan-
12140 guages*, pp. 69–123. Springer.
- 12141 Joshi, A. K., K. V. Shanker, and D. Weir (1991). The convergence of mildly context-sensitive
12142 grammar formalisms. In *Foundational Issues in Natural Language Processing*. Cambridge
12143 MA: MIT Press.

- 12144 Jozefowicz, R., O. Vinyals, M. Schuster, N. Shazeer, and Y. Wu (2016). Exploring the limits
12145 of language modeling. *arXiv preprint arXiv:1602.02410*.
- 12146 Jozefowicz, R., W. Zaremba, and I. Sutskever (2015). An empirical exploration of recurrent
12147 network architectures. In *Proceedings of the International Conference on Machine Learning*
12148 (*ICML*), pp. 2342–2350.
- 12149 Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambigua-
12150 tion. *Cognitive Science* 20(2), 137–194.
- 12151 Jurafsky, D. and J. H. Martin (2009). *Speech and Language Processing* (Second ed.). Prentice
12152 Hall.
- 12153 Jurafsky, D. and J. H. Martin (2018). *Speech and Language Processing* (Third ed.). Prentice
12154 Hall.
- 12155 Kadlec, R., M. Schmid, O. Bajgar, and J. Kleindienst (2016). Text understanding with
12156 the attention sum reader network. In *Proceedings of the Association for Computational*
12157 *Linguistics (ACL)*, pp. 908–918.
- 12158 Kalchbrenner, N. and P. Blunsom (2013). Recurrent convolutional neural networks for dis-
12159 course compositionality. In *Proceedings of the Workshop on Continuous Vector Space Models*
12160 and their Compositionality, pp. 119–126. Association for Computational Linguistics.
- 12161 Kalchbrenner, N., L. Espeholt, K. Simonyan, A. v. d. Oord, A. Graves, and K. Kavukcuoglu
12162 (2016). Neural machine translation in linear time. *arXiv preprint arXiv:1610.10099*.
- 12163 Kalchbrenner, N., E. Grefenstette, and P. Blunsom (2014). A convolutional neural network
12164 for modelling sentences. In *Proceedings of the Association for Computational Linguistics*
12165 (*ACL*), pp. 655–665.
- 12166 Karlsson, F. (2007). Constraints on multiple center-embedding of clauses. *Journal of Lin-*
12167 *guistics* 43(2), 365–392.
- 12168 Kate, R. J., Y. W. Wong, and R. J. Mooney (2005). Learning to transform natural to formal
12169 languages. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*.
- 12170 Kawaguchi, K., L. P. Kaelbling, and Y. Bengio (2017). Generalization in deep learning.
12171 *arXiv preprint arXiv:1710.05468*.
- 12172 Kehler, A. (2007). Rethinking the SMASH approach to pronoun interpretation. In *Interdis-*
12173 *ciplinary perspectives on reference processing*, New Directions in Cognitive Science Series,
12174 pp. 95–122. Oxford University Press.
- 12175 Kibble, R. and R. Power (2004). Optimizing referential coherence in text generation. *Com-*
12176 *putational Linguistics* 30(4), 401–416.

- 12177 Kilgarriff, A. (1997). I don't believe in word senses. *Computers and the Humanities* 31(2),
12178 91–113.
- 12179 Kilgarriff, A. and G. Grefenstette (2003). Introduction to the special issue on the web as
12180 corpus. *Computational linguistics* 29(3), 333–347.
- 12181 Kim, M.-J. (2002). Does Korean have adjectives? *MIT Working Papers in Linguistics* 43,
12182 71–89.
- 12183 Kim, S.-M. and E. Hovy (2006, July). Extracting opinions, opinion holders, and topics
12184 expressed in online news media text. In *Proceedings of the Workshop on Sentiment and*
12185 *Subjectivity in Text*, pp. 1–8. Association for Computational Linguistics.
- 12186 Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings*
12187 *of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 1746–1751.
- 12188 Kim, Y., C. Denton, L. Hoang, and A. M. Rush (2017). Structured attention networks. In
12189 *Proceedings of the International Conference on Learning Representations (ICLR)*.
- 12190 Kim, Y., Y. Jernite, D. Sontag, and A. M. Rush (2016). Character-aware neural language
12191 models. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*.
- 12192 Kingma, D. and J. Ba (2014). Adam: A method for stochastic optimization. *arXiv preprint*
12193 *arXiv:1412.6980*.
- 12194 Kiperwasser, E. and Y. Goldberg (2016). Simple and accurate dependency parsing using
12195 bidirectional lstm feature representations. *Transactions of the Association for Compu-*
12196 *tational Linguistics* 4, 313–327.
- 12197 Kipper-Schuler, K. (2005). *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph. D.
12198 thesis, Computer and Information Science, University of Pennsylvania.
- 12199 Kiros, R., R. Salakhutdinov, and R. Zemel (2014). Multimodal neural language models. In
12200 *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 595–603.
- 12201 Kiros, R., Y. Zhu, R. Salakhutdinov, R. S. Zemel, A. Torralba, R. Urtasun, and S. Fidler
12202 (2015). Skip-thought vectors. In *Neural Information Processing Systems (NIPS)*.
- 12203 Klein, D. and C. D. Manning (2003). Accurate unlexicalized parsing. In *Proceedings of the*
12204 *Association for Computational Linguistics (ACL)*, pp. 423–430.
- 12205 Klein, D. and C. D. Manning (2004). Corpus-based induction of syntactic structure: Mod-
12206 els of dependency and constituency. In *Proceedings of the Association for Computational*
12207 *Linguistics (ACL)*.

- 12208 Klein, G., Y. Kim, Y. Deng, J. Senellart, and A. M. Rush (2017). OpenNMT: Open-source
12209 toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*.
- 12210 Klementiev, A., I. Titov, and B. Bhattacharyya (2012). Inducing crosslingual distributed repre-
12211 sentations of words. In *Proceedings of the International Conference on Computational Lin-*
12212 *guistics (COLING)*, pp. 1459–1474.
- 12213 Klenner, M. (2007). Enforcing consistency on coreference sets. In *Recent Advances in Natu-*
12214 *ral Language Processing (RANLP)*, pp. 323–328.
- 12215 Knight, K. (1999). Decoding complexity in word-replacement translation models. *Compu-*
12216 *tational Linguistics* 25(4), 607–615.
- 12217 Knight, K. and J. Graehl (1998). Machine transliteration. *Computational Linguistics* 24(4),
12218 599–612.
- 12219 Knight, K. and D. Marcu (2000). Statistics-based summarization-step one: Sentence com-
12220 pression. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, pp.
12221 703–710.
- 12222 Knight, K. and J. May (2009). Applications of weighted automata in natural language
12223 processing. In *Handbook of Weighted Automata*, pp. 571–596. Springer.
- 12224 Knott, A. (1996). *A data-driven methodology for motivating a set of coherence relations*. Ph. D.
12225 thesis, The University of Edinburgh.
- 12226 Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT*
12227 *summit*, Volume 5, pp. 79–86.
- 12228 Koehn, P. (2009). *Statistical machine translation*. Cambridge University Press.
- 12229 Koehn, P. (2017). Neural machine translation. *arXiv preprint arXiv:1709.07809*.
- 12230 Konstas, I. and M. Lapata (2013). A global model for concept-to-text generation. *Journal*
12231 *of Artificial Intelligence Research* 48, 305–346.
- 12232 Koo, T., X. Carreras, and M. Collins (2008). Simple semi-supervised dependency parsing.
12233 In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 595–603.
- 12234 Koo, T. and M. Collins (2005). Hidden-variable models for discriminative reranking. In
12235 *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 507–514.
- 12236 Koo, T. and M. Collins (2010). Efficient third-order dependency parsers. In *Proceedings of*
12237 *the Association for Computational Linguistics (ACL)*.

- 12238 Koo, T., A. Globerson, X. Carreras, and M. Collins (2007). Structured prediction models
12239 via the matrix-tree theorem. In *Proceedings of Empirical Methods for Natural Language
12240 Processing (EMNLP)*, pp. 141–150.
- 12241 Kovach, B. and T. Rosenstiel (2014). *The elements of journalism: What newspeople should know
12242 and the public should expect*. Three Rivers Press.
- 12243 Krishnamurthy, J. (2016). Probabilistic models for learning a semantic parser lexicon. In
12244 *Proceedings of the North American Chapter of the Association for Computational Linguistics
12245 (NAACL)*, pp. 606–616.
- 12246 Krishnamurthy, J. and T. M. Mitchell (2012). Weakly supervised training of semantic
12247 parsers. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*,
12248 pp. 754–765.
- 12249 Krizhevsky, A., I. Sutskever, and G. E. Hinton (2012). Imagenet classification with deep
12250 convolutional neural networks. In *Neural Information Processing Systems (NIPS)*, pp.
12251 1097–1105.
- 12252 Kübler, S., R. McDonald, and J. Nivre (2009). Dependency parsing. *Synthesis Lectures on
12253 Human Language Technologies* 1(1), 1–127.
- 12254 Kuhlmann, M. and J. Nivre (2010). Transition-based techniques for non-projective depen-
12255 dency parsing. *Northern European Journal of Language Technology (NEJLT)* 2(1), 1–19.
- 12256 Kummerfeld, J. K., T. Berg-Kirkpatrick, and D. Klein (2015). An empirical analysis of op-
12257 timization for max-margin NLP. In *Proceedings of Empirical Methods for Natural Language
12258 Processing (EMNLP)*.
- 12259 Kwiatkowski, T., S. Goldwater, L. Zettlemoyer, and M. Steedman (2012). A probabilistic
12260 model of syntactic and semantic acquisition from child-directed utterances and their
12261 meanings. In *Proceedings of the European Chapter of the Association for Computational Lin-
12262 guistics (EACL)*, pp. 234–244.
- 12263 Lafferty, J., A. McCallum, and F. Pereira (2001). Conditional random fields: Probabilistic
12264 models for segmenting and labeling sequence data. In *Proceedings of the International
12265 Conference on Machine Learning (ICML)*.
- 12266 Lakoff, G. (1973). Hedges: A study in meaning criteria and the logic of fuzzy concepts.
12267 *Journal of philosophical logic* 2(4), 458–508.
- 12268 Lample, G., M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer (2016). Neural
12269 architectures for named entity recognition. In *Proceedings of the North American Chapter
12270 of the Association for Computational Linguistics (NAACL)*, pp. 260–270.

- 12271 Langkilde, I. and K. Knight (1998). Generation that exploits corpus-based statistical
12272 knowledge. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 704–
12273 710.
- 12274 Lapata, M. (2003). Probabilistic text structuring: Experiments with sentence ordering. In
12275 *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 545–552.
- 12276 Lappin, S. and H. J. Leass (1994). An algorithm for pronominal anaphora resolution.
12277 *Computational linguistics* 20(4), 535–561.
- 12278 Lari, K. and S. J. Young (1990). The estimation of stochastic context-free grammars using
12279 the inside-outside algorithm. *Computer speech & language* 4(1), 35–56.
- 12280 Lascarides, A. and N. Asher (2007). Segmented discourse representation theory: Dynamic
12281 semantics with discourse structure. In *Computing meaning*, pp. 87–124. Springer.
- 12282 Law, E. and L. v. Ahn (2011). Human computation. *Synthesis Lectures on Artificial Intelli-*
12283 *gence and Machine Learning* 5(3), 1–121.
- 12284 Lebret, R., D. Grangier, and M. Auli (2016). Neural text generation from structured data
12285 with application to the biography domain. In *Proceedings of Empirical Methods for Natural*
12286 *Language Processing (EMNLP)*, pp. 1203–1213.
- 12287 LeCun, Y., L. Bottou, G. B. Orr, and K.-R. Müller (1998). Efficient backprop. In *Neural*
12288 *networks: Tricks of the trade*, pp. 9–50. Springer.
- 12289 Lee, C. M. and S. S. Narayanan (2005). Toward detecting emotions in spoken dialogs.
12290 *IEEE Transactions on speech and audio processing* 13(2), 293–303.
- 12291 Lee, H., A. Chang, Y. Peirsman, N. Chambers, M. Surdeanu, and D. Jurafsky (2013). De-
12292 terministic coreference resolution based on entity-centric, precision-ranked rules. *Com-*
12293 *putational Linguistics* 39(4), 885–916.
- 12294 Lee, H., Y. Peirsman, A. Chang, N. Chambers, M. Surdeanu, and D. Jurafsky (2011). Stan-
12295 ford’s multi-pass sieve coreference resolution system at the conll-2011 shared task. In
12296 *Proceedings of the Conference on Natural Language Learning (CoNLL)*, pp. 28–34.
- 12297 Lee, K., L. He, M. Lewis, and L. Zettlemoyer (2017). End-to-end neural coreference reso-
12298 lution. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.
- 12299 Lee, K., L. He, and L. Zettlemoyer (2018). Higher-order coreference resolution with coarse-
12300 to-fine inference. In *Proceedings of the North American Chapter of the Association for Com-*
12301 *putational Linguistics (NAACL)*, pp. 687–692.
- 12302 Lenat, D. B., R. V. Guha, K. Pittman, D. Pratt, and M. Shepherd (1990). Cyc: toward
12303 programs with common sense. *Communications of the ACM* 33(8), 30–49.

- 12304 Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries:
12305 how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, pp. 24–26. ACM.
- 12307 Levesque, H. J., E. Davis, and L. Morgenstern (2011). The Winograd schema challenge. In
12308 *AAAI spring symposium: Logical formalizations of commonsense reasoning*.
- 12309 Levin, E., R. Pieraccini, and W. Eckert (1998). Using markov decision process for learning
12310 dialogue strategies. In *Proceedings of the International Conference on Acoustics, Speech and*
12311 *Signal Processing*, Volume 1, pp. 201–204. IEEE.
- 12312 Levy, O. and Y. Goldberg (2014). Dependency-based word embeddings. In *Proceedings of*
12313 *the Association for Computational Linguistics (ACL)*, pp. 302–308.
- 12314 Levy, O., Y. Goldberg, and I. Dagan (2015). Improving distributional similarity with
12315 lessons learned from word embeddings. *Transactions of the Association for Computational*
12316 *Linguistics* 3, 211–225.
- 12317 Levy, R. and C. Manning (2009). An informal introduction to computational semantics.
12318 http://idiom.ucsd.edu/~rlevy/teaching/winter2009/ligncse256/lectures/lecture_14_compositional_semantics.pdf.
- 12320 Lewis, M. and M. Steedman (2013). Combined distributional and logical semantics. *Transactions of the Association for Computational Linguistics* 1, 179–192.
- 12322 Lewis II, P. M. and R. E. Stearns (1968). Syntax-directed transduction. *Journal of the*
12323 *ACM* 15(3), 465–488.
- 12324 Li, J. and D. Jurafsky (2015). Do multi-sense embeddings improve natural language
12325 understanding? In *Proceedings of Empirical Methods for Natural Language Processing*
12326 (*EMNLP*), pp. 1722–1732.
- 12327 Li, J. and D. Jurafsky (2017). Neural net models of open-domain discourse coherence. In
12328 *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 198–209.
- 12329 Li, J., R. Li, and E. Hovy (2014). Recursive deep models for discourse parsing. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.
- 12331 Li, J., M.-T. Luong, and D. Jurafsky (2015). A hierarchical neural autoencoder for para-
12332 graphs and documents. In *Proceedings of Empirical Methods for Natural Language Process-
12333 ing (EMNLP)*.
- 12334 Li, J., T. Luong, D. Jurafsky, and E. Hovy (2015). When are tree structures necessary
12335 for deep learning of representations? In *Proceedings of Empirical Methods for Natural
12336 Language Processing (EMNLP)*, pp. 2304–2314.

- 12337 Li, J., W. Monroe, A. Ritter, D. Jurafsky, M. Galley, and J. Gao (2016). Deep reinforcement
12338 learning for dialogue generation. In *Proceedings of Empirical Methods for Natural Language
12339 Processing (EMNLP)*, pp. 1192–1202.
- 12340 Li, Q., S. Anzaroot, W.-P. Lin, X. Li, and H. Ji (2011). Joint inference for cross-document
12341 information extraction. In *Proceedings of the International Conference on Information and
12342 Knowledge Management (CIKM)*, pp. 2225–2228.
- 12343 Li, Q., H. Ji, and L. Huang (2013). Joint event extraction via structured prediction with
12344 global features. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp.
12345 73–82.
- 12346 Liang, P. (2005). Semi-supervised learning for natural language. Master’s thesis, Mas-
12347 sachusetts Institute of Technology.
- 12348 Liang, P., A. Bouchard-Côté, D. Klein, and B. Taskar (2006). An end-to-end discrimina-
12349 tive approach to machine translation. In *Proceedings of the Association for Computational
12350 Linguistics (ACL)*, pp. 761–768.
- 12351 Liang, P., M. Jordan, and D. Klein (2009). Learning semantic correspondences with less
12352 supervision. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 91–
12353 99.
- 12354 Liang, P., M. I. Jordan, and D. Klein (2013). Learning dependency-based compositional
12355 semantics. *Computational Linguistics* 39(2), 389–446.
- 12356 Liang, P. and D. Klein (2009). Online EM for unsupervised models. In *Proceedings of
12357 the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp.
12358 611–619.
- 12359 Liang, P., S. Petrov, M. I. Jordan, and D. Klein (2007). The infinite PCFG using hierarchical
12360 Dirichlet processes. In *Proceedings of Empirical Methods for Natural Language Processing
12361 (EMNLP)*, pp. 688–697.
- 12362 Liang, P. and C. Potts (2015). Bringing machine learning and compositional semantics
12363 together. *Annual Review of Linguistics* 1(1), 355–376.
- 12364 Lieber, R. (2015). *Introducing morphology*. Cambridge University Press.
- 12365 Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the
12366 International Conference on Computational Linguistics (COLING)*, pp. 768–774.
- 12367 Lin, J. and C. Dyer (2010). Data-intensive text processing with mapreduce. *Synthesis
12368 Lectures on Human Language Technologies* 3(1), 1–177.

- 12369 Lin, Z., M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio (2017). A struc-
12370 tured self-attentive sentence embedding. In *Proceedings of the International Conference on*
12371 *Learning Representations (ICLR)*.
- 12372 Lin, Z., M.-Y. Kan, and H. T. Ng (2009). Recognizing implicit discourse relations in the
12373 penn discourse treebank. In *Proceedings of Empirical Methods for Natural Language Pro-*
12374 *cessing (EMNLP)*, pp. 343–351.
- 12375 Lin, Z., H. T. Ng, and M.-Y. Kan (2011). Automatically evaluating text coherence using
12376 discourse relations. In *Proceedings of the Association for Computational Linguistics (ACL)*,
12377 pp. 997–1006.
- 12378 Lin, Z., H. T. Ng, and M.-Y. Kan (2014). A pdtb-styled end-to-end discourse parser. *Natural*
12379 *Language Engineering* 20(2), 151–184.
- 12380 Ling, W., C. Dyer, A. Black, and I. Trancoso (2015). Two/too simple adaptations of
12381 word2vec for syntax problems. In *Proceedings of the North American Chapter of the As-*
12382 *sociation for Computational Linguistics (NAACL)*.
- 12383 Ling, W., T. Luís, L. Marujo, R. F. Astudillo, S. Amir, C. Dyer, A. W. Black, and I. Trancoso
12384 (2015). Finding function in form: Compositional character models for open vocabulary
12385 word representation. In *Proceedings of Empirical Methods for Natural Language Processing*
12386 (*EMNLP*).
- 12387 Ling, W., G. Xiang, C. Dyer, A. Black, and I. Trancoso (2013). Microblogs as parallel cor-
12388 pora. In *Proceedings of the Association for Computational Linguistics (ACL)*.
- 12389 Ling, X., S. Singh, and D. S. Weld (2015). Design challenges for entity linking. *Transactions*
12390 *of the Association for Computational Linguistics* 3, 315–328.
- 12391 Linguistic Data Consortium (2005). ACE (automatic content extraction) English annota-
12392 tion guidelines for relations. Technical Report Version 5.8.3, Linguistic Data Consor-
12393 tium.
- 12394 Liu, B. (2015). *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge
12395 University Press.
- 12396 Liu, D. C. and J. Nocedal (1989). On the limited memory BFGS method for large scale
12397 optimization. *Mathematical programming* 45(1-3), 503–528.
- 12398 Liu, Y., Q. Liu, and S. Lin (2006). Tree-to-string alignment template for statistical machine
12399 translation. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 609–
12400 616.

- 12401 Loper, E. and S. Bird (2002). NLTK: The natural language toolkit. In *Proceedings of the*
12402 *Workshop on Effective tools and methodologies for teaching natural language processing and*
12403 *computational linguistics*, pp. 63–70. Association for Computational Linguistics.
- 12404 Louis, A., A. Joshi, and A. Nenkova (2010). Discourse indicators for content selection
12405 in summarization. In *Proceedings of the Special Interest Group on Discourse and Dialogue*
12406 (*SIGDIAL*), pp. 147–156.
- 12407 Louis, A. and A. Nenkova (2013). What makes writing great? first experiments on article
12408 quality prediction in the science journalism domain. *Transactions of the Association for*
12409 *Computational Linguistics* 1, 341–352.
- 12410 Loveland, D. W. (2016). *Automated Theorem Proving: a logical basis*. Elsevier.
- 12411 Lowe, R., N. Pow, I. V. Serban, and J. Pineau (2015). The Ubuntu Dialogue Corpus: A
12412 large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings*
12413 *of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*.
- 12414 Luo, X. (2005). On coreference resolution performance metrics. In *Proceedings of Empirical*
12415 *Methods for Natural Language Processing (EMNLP)*, pp. 25–32.
- 12416 Luo, X., A. Ittycheriah, H. Jing, N. Kambhatla, and S. Roukos (2004). A mention-
12417 synchronous coreference resolution algorithm based on the bell tree. In *Proceedings*
12418 *of the Association for Computational Linguistics (ACL)*.
- 12419 Luong, M.-T., R. Socher, and C. D. Manning (2013). Better word representations with
12420 recursive neural networks for morphology.
- 12421 Luong, T., H. Pham, and C. D. Manning (2015). Effective approaches to attention-based
12422 neural machine translation. In *Proceedings of Empirical Methods for Natural Language*
12423 *Processing (EMNLP)*, pp. 1412–1421.
- 12424 Luong, T., I. Sutskever, Q. Le, O. Vinyals, and W. Zaremba (2015). Addressing the rare
12425 word problem in neural machine translation. In *Proceedings of the Association for Compu-*
12426 *tational Linguistics (ACL)*, pp. 11–19.
- 12427 Maas, A. L., A. Y. Hannun, and A. Y. Ng (2013). Rectifier nonlinearities improve neu-
12428 *ral network acoustic models*. In *Proceedings of the International Conference on Machine*
12429 *Learning (ICML)*.
- 12430 Magerman, D. M. (1995). Statistical decision-tree models for parsing. In *Proceedings of the*
12431 *Association for Computational Linguistics (ACL)*, pp. 276–283.
- 12432 Mairesse, F. and M. A. Walker (2011). Controlling user perceptions of linguistic style:
12433 Trainable generation of personality traits. *Computational Linguistics* 37(3), 455–488.

- 12434 Mani, I., M. Verhagen, B. Wellner, C. M. Lee, and J. Pustejovsky (2006). Machine learning
 12435 of temporal relations. In *Proceedings of the Association for Computational Linguistics (ACL)*,
 12436 pp. 753–760.
- 12437 Mann, W. C. and S. A. Thompson (1988). Rhetorical structure theory: Toward a functional
 12438 theory of text organization. *Text* 8(3), 243–281.
- 12439 Manning, C. D. (2015). Last words: Computational linguistics and deep learning. *Computa-
 12440 tional Linguistics* 41(4), 701–707.
- 12441 Manning, C. D., P. Raghavan, H. Schütze, et al. (2008). *Introduction to information retrieval*,
 12442 Volume 1. Cambridge university press.
- 12443 Manning, C. D. and H. Schütze (1999). *Foundations of Statistical Natural Language Process-
 12444 ing*. Cambridge, Massachusetts: MIT press.
- 12445 Marcu, D. (1996). Building up rhetorical structure trees. In *Proceedings of the National
 12446 Conference on Artificial Intelligence*, pp. 1069–1074.
- 12447 Marcu, D. (1997a). From discourse structures to text summaries. In *Proceedings of the
 12448 workshop on Intelligent Scalable Text Summarization*.
- 12449 Marcu, D. (1997b). From local to global coherence: A bottom-up approach to text plan-
 12450 ning. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, pp. 629–635.
- 12451 Marcus, M. P., M. A. Marcinkiewicz, and B. Santorini (1993). Building a large annotated
 12452 corpus of English: The Penn Treebank. *Computational Linguistics* 19(2), 313–330.
- 12453 Maron, O. and T. Lozano-Pérez (1998). A framework for multiple-instance learning. In
 12454 *Neural Information Processing Systems (NIPS)*, pp. 570–576.
- 12455 Márquez, G. G. (1970). *One Hundred Years of Solitude*. Harper & Row. English translation
 12456 by Gregory Rabassa.
- 12457 Martins, A. F. T., N. A. Smith, and E. P. Xing (2009). Concise integer linear programming
 12458 formulations for dependency parsing. In *Proceedings of the Association for Computational
 12459 Linguistics (ACL)*, pp. 342–350.
- 12460 Martins, A. F. T., N. A. Smith, E. P. Xing, P. M. Q. Aguiar, and M. A. T. Figueiredo (2010).
 12461 Turbo parsers: Dependency parsing by approximate variational inference. In *Proceed-
 12462 ings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 34–44.
- 12463 Matsuzaki, T., Y. Miyao, and J. Tsujii (2005). Probabilistic cfg with latent annotations. In
 12464 *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 75–82.

- 12465 Matthiessen, C. and J. A. Bateman (1991). *Text generation and systemic-functional linguistics: experiences from English and Japanese*. Pinter Publishers.
- 12466
- 12467 McCallum, A. and W. Li (2003). Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 188–191.
- 12468
- 12469
- 12470
- 12471 McCallum, A. and B. Wellner (2004). Conditional models of identity uncertainty with application to noun coreference. In *Neural Information Processing Systems (NIPS)*, pp. 905–912.
- 12472
- 12473
- 12474 McDonald, R., K. Crammer, and F. Pereira (2005). Online large-margin training of dependency parsers. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 91–98.
- 12475
- 12476
- 12477 McDonald, R., K. Hannan, T. Neylon, M. Wells, and J. Reynar (2007). Structured models for fine-to-coarse sentiment analysis. In *Proceedings of the Association for Computational Linguistics (ACL)*.
- 12478
- 12479
- 12480 McDonald, R. and F. Pereira (2006). Online learning of approximate dependency parsing algorithms. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*.
- 12481
- 12482
- 12483 McKeown, K. (1992). *Text generation*. Cambridge University Press.
- 12484
- 12485
- 12486
- 12487 McKeown, K., S. Rosenthal, K. Thadani, and C. Moore (2010). Time-efficient creation of an accurate sentence fusion corpus. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 317–320.
- 12488
- 12489
- 12490
- 12491 McNamee, P. and H. T. Dang (2009). Overview of the tac 2009 knowledge base population track. In *Text Analysis Conference (TAC)*, Volume 17, pp. 111–113.
- 12492
- 12493 Medlock, B. and T. Briscoe (2007). Weakly supervised learning for hedge classification in scientific literature. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 992–999.
- 12494
- 12495
- 12496 Mei, H., M. Bansal, and M. R. Walter (2016). What to talk about and how? selective generation using lstms with coarse-to-fine alignment. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 720–730.
- 12497
- 12498

- 12499 Merity, S., N. S. Keskar, and R. Socher (2018). Regularizing and optimizing LSTM language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- 12502 Merity, S., C. Xiong, J. Bradbury, and R. Socher (2017). Pointer sentinel mixture models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- 12504 Messud, C. (2014, June). A new ‘l’étranger’. *New York Review of Books*.
- 12505 Miao, Y. and P. Blunsom (2016). Language as a latent variable: Discrete generative models for sentence compression. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 319–328.
- 12508 Miao, Y., L. Yu, and P. Blunsom (2016). Neural variational inference for text processing. In *Proceedings of the International Conference on Machine Learning (ICML)*.
- 12510 Mihalcea, R., T. A. Chklovski, and A. Kilgarriff (2004). The SENSEVAL-3 English lexical sample task. In *Proceedings of SENSEVAL-3*, Barcelona, Spain, pp. 25–28. Association for Computational Linguistics.
- 12513 Mihalcea, R. and D. Radev (2011). *Graph-based natural language processing and information retrieval*. Cambridge University Press.
- 12515 Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013). Efficient estimation of word representations in vector space. In *Proceedings of International Conference on Learning Representations*.
- 12518 Mikolov, T., A. Deoras, D. Povey, L. Burget, and J. Cernocký (2011). Strategies for training large scale neural network language models. In *Proceedings of the Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 196–201.
- 12521 Mikolov, T., M. Karafiat, L. Burget, J. Cernocký, and S. Khudanpur (2010). Recurrent neural network based language model. In *INTERSPEECH*, pp. 1045–1048.
- 12523 Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pp. 3111–3119.
- 12526 Mikolov, T., W.-t. Yih, and G. Zweig (2013). Linguistic regularities in continuous space word representations. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 746–751.
- 12529 Mikolov, T. and G. Zweig. Context dependent recurrent neural network language model. In *Proceedings of Spoken Language Technology (SLT)*, pp. 234–239.

- 12531 Miller, G. A., G. A. Heise, and W. Lichten (1951). The intelligibility of speech as a function
12532 of the context of the test materials. *Journal of experimental psychology* 41(5), 329.
- 12533 Miller, M., C. Sathi, D. Wiesenthal, J. Leskovec, and C. Potts (2011). Sentiment flow
12534 through hyperlink networks. In *Proceedings of the International Conference on Web and*
12535 *Social Media (ICWSM)*.
- 12536 Miller, S., J. Guinness, and A. Zamanian (2004). Name tagging with word clusters and
12537 discriminative training. In *Proceedings of the North American Chapter of the Association for*
12538 *Computational Linguistics (NAACL)*, pp. 337–342.
- 12539 Milne, D. and I. H. Witten (2008). Learning to link with wikipedia. In *Proceedings of the*
12540 *International Conference on Information and Knowledge Management (CIKM)*, pp. 509–518.
- 12541 Miltsakaki, E. and K. Kukich (2004). Evaluation of text coherence for electronic essay
12542 scoring systems. *Natural Language Engineering* 10(1), 25–55.
- 12543 Minka, T. P. (1999). From hidden markov models to linear dynamical systems. Tech. Rep.
12544 531, Vision and Modeling Group of Media Lab, MIT.
- 12545 Minsky, M. (1974). A framework for representing knowledge. Technical Report 306, MIT
12546 AI Laboratory.
- 12547 Minsky, M. and S. Papert (1969). *Perceptrons*. MIT press.
- 12548 Mintz, M., S. Bills, R. Snow, and D. Jurafsky (2009). Distant supervision for relation extrac-
12549 tion without labeled data. In *Proceedings of the Association for Computational Linguistics*
12550 (*ACL*), pp. 1003–1011.
- 12551 Mirza, P., R. Sprugnoli, S. Tonelli, and M. Speranza (2014). Annotating causality in the
12552 TempEval-3 corpus. In *Proceedings of the EACL 2014 Workshop on Computational Ap-*
12553 *proaches to Causality in Language (CAtoCL)*, pp. 10–19.
- 12554 Misra, D. K. and Y. Artzi (2016). Neural shift-reduce ccg semantic parsing. In *Proceedings*
12555 *of Empirical Methods for Natural Language Processing (EMNLP)*.
- 12556 Mitchell, J. and M. Lapata (2010). Composition in distributional models of semantics.
12557 *Cognitive Science* 34(8), 1388–1429.
- 12558 Miwa, M. and M. Bansal (2016). End-to-end relation extraction using lstms on sequences
12559 and tree structures. In *Proceedings of the Association for Computational Linguistics (ACL)*,
12560 pp. 1105–1116.
- 12561 Mnih, A. and G. Hinton (2007). Three new graphical models for statistical language mod-
12562 ellng. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 641–
12563 648.

- 12564 Mnih, A. and G. E. Hinton (2008). A scalable hierarchical distributed language model. In
 12565 *Neural Information Processing Systems (NIPS)*, pp. 1081–1088.
- 12566 Mnih, A. and Y. W. Teh (2012). A fast and simple algorithm for training neural probabilis-
 12567 tic language models. In *Proceedings of the International Conference on Machine Learning*
 12568 (*ICML*).
- 12569 Mohammad, S. M. and P. D. Turney (2013). Crowdsourcing a word–emotion association
 12570 lexicon. *Computational Intelligence* 29(3), 436–465.
- 12571 Mohri, M., F. Pereira, and M. Riley (2002). Weighted finite-state transducers in speech
 12572 recognition. *Computer Speech & Language* 16(1), 69–88.
- 12573 Mohri, M., A. Rostamizadeh, and A. Talwalkar (2012). *Foundations of machine learning*.
 12574 MIT press.
- 12575 Montague, R. (1973). The proper treatment of quantification in ordinary english. In *Ap-
 12576 proaches to natural language*, pp. 221–242. Springer.
- 12577 Moore, J. D. and C. L. Paris (1993). Planning text for advisory dialogues: Capturing inten-
 12578 tional and rhetorical information. *Computational Linguistics* 19(4), 651–694.
- 12579 Morante, R. and E. Blanco (2012). *SEM 2012 shared task: Resolving the scope and focus of
 12580 negation. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*,
 12581 pp. 265–274. Association for Computational Linguistics.
- 12582 Morante, R. and W. Daelemans (2009). Learning the scope of hedge cues in biomedical
 12583 texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language
 12584 Processing*, pp. 28–36. Association for Computational Linguistics.
- 12585 Morante, R. and C. Sporleder (2012). Modality and negation: An introduction to the
 12586 special issue. *Computational linguistics* 38(2), 223–260.
- 12587 Mostafazadeh, N., A. Grelish, N. Chambers, J. Allen, and L. Vanderwende (2016). CaTeRS:
 12588 Causal and temporal relation scheme for semantic annotation of event struc-
 12589 tures. In *Proceedings of the Fourth Workshop on Events*, pp. 51–61. Association for Com-
 12590 putational Linguistics.
- 12591 Mueller, T., H. Schmid, and H. Schütze (2013). Efficient higher-order CRFs for morpholog-
 12592 ical tagging. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*,
 12593 pp. 322–332.
- 12594 Müller, C. and M. Strube (2006). Multi-level annotation of linguistic data with mmax2.
 12595 *Corpus technology and language pedagogy: New resources, new tools, new methods* 3, 197–
 12596 214.

- 12597 Muralidharan, A. and M. A. Hearst (2013). Supporting exploratory text analysis in literature study. *Literary and linguistic computing* 28(2), 283–295.
- 12598
- 12599 Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press.
- 12600 Nakagawa, T., K. Inui, and S. Kurohashi (2010). Dependency tree-based sentiment classification using CRFs with hidden variables. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 786–794.
- 12601
- 12602
- 12603 Nakazawa, T., M. Yaguchi, K. Uchimoto, M. Utiyama, E. Sumita, S. Kurohashi, and H. Isahara (2016). ASPEC: Asian scientific paper excerpt corpus. In *Proceedings of the Language Resources and Evaluation Conference*, pp. 2204–2208.
- 12604
- 12605
- 12606 Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys* 41(2), 10.
- 12607
- 12608 Neal, R. M. and G. E. Hinton (1998). A view of the em algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pp. 355–368. Springer.
- 12609
- 12610 Nenkova, A. and K. McKeown (2012). A survey of text summarization techniques. In *Mining text data*, pp. 43–76. Springer.
- 12611
- 12612 Neubig, G. (2017). Neural machine translation and sequence-to-sequence models: A tutorial. *arXiv preprint arXiv:1703.01619*.
- 12613
- 12614 Neubig, G., C. Dyer, Y. Goldberg, A. Matthews, W. Ammar, A. Anastasopoulos, M. Balles-
12615 teros, D. Chiang, D. Clothiaux, T. Cohn, K. Duh, M. Faruqui, C. Gan, D. Garrette,
12616 Y. Ji, L. Kong, A. Kuncoro, G. Kumar, C. Malaviya, P. Michel, Y. Oda, M. Richardson,
12617 N. Saphra, S. Swayamdipta, and P. Yin (2017). Dynet: The dynamic neural network
12618 toolkit.
- 12619 Neubig, G., Y. Goldberg, and C. Dyer (2017). On-the-fly operation batching in dynamic
12620 computation graphs. In *Neural Information Processing Systems (NIPS)*.
- 12621 Neubig, G., M. Sperber, X. Wang, M. Felix, A. Matthews, S. Padmanabhan, Y. Qi, D. S.
12622 Sachan, P. Arthur, P. Godard, J. Hewitt, R. Riad, and L. Wang (2018). XNMT: The ex-
12623 tensible neural machine translation toolkit. In *Conference of the Association for Machine
12624 Translation in the Americas (AMTA)*.
- 12625 Neuhaus, P. and N. Bröker (1997). The complexity of recognition of linguistically ade-
12626 quate dependency grammars. In *Proceedings of the European Chapter of the Association for
12627 Computational Linguistics (EACL)*, pp. 337–343.
- 12628 Newman, D., C. Chemudugunta, and P. Smyth (2006). Statistical entity-topic models. In
12629 *Proceedings of Knowledge Discovery and Data Mining (KDD)*, pp. 680–686.

- 12630 Ng, V. (2010). Supervised noun phrase coreference research: The first fifteen years. In
 12631 *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 1396–1411.
- 12632 Nguyen, D. and A. S. Dogruöz (2013). Word level language identification in online multi-
 12633 lingual communication. In *Proceedings of Empirical Methods for Natural Language Process-
 12634 ing (EMNLP)*.
- 12635 Nguyen, D. T. and S. Joty (2017). A neural local coherence model. In *Proceedings of the
 12636 Association for Computational Linguistics (ACL)*, pp. 1320–1330.
- 12637 Nigam, K., A. K. McCallum, S. Thrun, and T. Mitchell (2000). Text classification from
 12638 labeled and unlabeled documents using em. *Machine learning* 39(2-3), 103–134.
- 12639 Nirenburg, S. and Y. Wilks (2001). What’s in a symbol: ontology, representation and lan-
 12640 guage. *Journal of Experimental & Theoretical Artificial Intelligence* 13(1), 9–23.
- 12641 Nivre, J. (2008). Algorithms for deterministic incremental dependency parsing. *Compu-
 12642 tational Linguistics* 34(4), 513–553.
- 12643 Nivre, J., M.-C. de Marneffe, F. Ginter, Y. Goldberg, J. Hajic, C. D. Manning, R. McDonald,
 12644 S. Petrov, S. Pyysalo, N. Silveira, R. Tsarfaty, and D. Zeman (2016). Universal depen-
 12645 dencies v1: A multilingual treebank collection. In *Proceedings of the Language Resources
 12646 and Evaluation Conference*.
- 12647 Nivre, J. and J. Nilsson (2005). Pseudo-projective dependency parsing. In *Proceedings of
 12648 the Association for Computational Linguistics (ACL)*, pp. 99–106.
- 12649 Novikoff, A. B. (1962). On convergence proofs on perceptrons. In *Proceedings of the Sym-
 12650 posium on the Mathematical Theory of Automata*, Volume 12, pp. 615—622.
- 12651 Och, F. J. and H. Ney (2003). A systematic comparison of various statistical alignment
 12652 models. *Computational linguistics* 29(1), 19–51.
- 12653 O’Connor, B., M. Krieger, and D. Ahn (2010). Tweetmotif: Exploratory search and topic
 12654 summarization for twitter. In *Proceedings of the International Conference on Web and Social
 12655 Media (ICWSM)*, pp. 384–385.
- 12656 Oepen, S., M. Kuhlmann, Y. Miyao, D. Zeman, D. Flickinger, J. Hajic, A. Ivanova, and
 12657 Y. Zhang (2014). Semeval 2014 task 8: Broad-coverage semantic dependency parsing.
 12658 In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pp.
 12659 63–72.
- 12660 Oflazer, K. and İ. Kuruöz (1994). Tagging and morphological disambiguation of turkish
 12661 text. In *Proceedings of the fourth conference on Applied natural language processing*, pp. 144–
 12662 149.

- 12663 Ohta, T., Y. Tateisi, and J.-D. Kim (2002). The genia corpus: An annotated research abstract
12664 corpus in molecular biology domain. In *Proceedings of the second international conference*
12665 on *Human Language Technology Research*, pp. 82–86. Morgan Kaufmann Publishers Inc.
- 12666 Onishi, T., H. Wang, M. Bansal, K. Gimpel, and D. McAllester (2016). Who did what: A
12667 large-scale person-centered cloze dataset. In *Proceedings of Empirical Methods for Natural*
12668 *Language Processing (EMNLP)*, pp. 2230–2235.
- 12669 Owoputi, O., B. O'Connor, C. Dyer, K. Gimpel, N. Schneider, and N. A. Smith (2013).
12670 Improved part-of-speech tagging for online conversational text with word clusters. In
12671 *Proceedings of the North American Chapter of the Association for Computational Linguistics*
12672 (*NAACL*), pp. 380–390.
- 12673 Packard, W., E. M. Bender, J. Read, S. Oepen, and R. Dridan (2014). Simple negation
12674 scope resolution through deep parsing: A semantic solution to a semantic problem. In
12675 *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 69–78.
- 12676 Paice, C. D. (1990). Another stemmer. In *ACM SIGIR Forum*, Volume 24, pp. 56–61.
- 12677 Pak, A. and P. Paroubek (2010). Twitter as a corpus for sentiment analysis and opinion
12678 mining. In *Proceedings of the Language Resources and Evaluation Conference*, pp. 1320–1326.
- 12679 Palmer, F. R. (2001). *Mood and modality*. Cambridge University Press.
- 12680 Palmer, M., D. Gildea, and P. Kingsbury (2005). The proposition bank: An annotated
12681 corpus of semantic roles. *Computational linguistics* 31(1), 71–106.
- 12682 Pan, S. J. and Q. Yang (2010). A survey on transfer learning. *IEEE Transactions on knowledge*
12683 *and data engineering* 22(10), 1345–1359.
- 12684 Pang, B. and L. Lee (2004). A sentimental education: Sentiment analysis using subjectivity
12685 summarization based on minimum cuts. In *Proceedings of the Association for Computational*
12686 *Linguistics (ACL)*, pp. 271–278.
- 12687 Pang, B. and L. Lee (2005). Seeing stars: Exploiting class relationships for sentiment cate-
12688 gorization with respect to rating scales. In *Proceedings of the Association for Computational*
12689 *Linguistics (ACL)*, pp. 115–124.
- 12690 Pang, B. and L. Lee (2008). Opinion mining and sentiment analysis. *Foundations and trends*
12691 in *information retrieval* 2(1-2), 1–135.
- 12692 Pang, B., L. Lee, and S. Vaithyanathan (2002). Thumbs up?: sentiment classification using
12693 machine learning techniques. In *Proceedings of Empirical Methods for Natural Language*
12694 *Processing (EMNLP)*, pp. 79–86.

- 12695 Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu (2002). Bleu: a method for automatic
12696 evaluation of machine translation. In *Proceedings of the Association for Computational
12697 Linguistics (ACL)*, pp. 311–318.
- 12698 Park, J. and C. Cardie (2012). Improving implicit discourse relation recognition through
12699 feature set optimization. In *Proceedings of the Special Interest Group on Discourse and Dia-
12700 logue (SIGDIAL)*, pp. 108–112.
- 12701 Parsons, T. (1990). *Events in the Semantics of English*, Volume 5. MIT Press.
- 12702 Pascanu, R., T. Mikolov, and Y. Bengio (2013). On the difficulty of training recurrent neural
12703 networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp.
12704 1310–1318.
- 12705 Paul, M., M. Federico, and S. Stüker (2010). Overview of the iwslt 2010 evaluation cam-
12706 paign. In *International Workshop on Spoken Language Translation (IWSLT) 2010*.
- 12707 Pedersen, T., S. Patwardhan, and J. Michelizzi (2004). Wordnet::similarity - measuring the
12708 relatedness of concepts. In *Proceedings of the North American Chapter of the Association for
12709 Computational Linguistics (NAACL)*, pp. 38–41.
- 12710 Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blon-
12711 del, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau,
12712 M. Brucher, M. Perrot, and E. Duchesnay (2011). Scikit-learn: Machine learning in
12713 Python. *Journal of Machine Learning Research* 12, 2825–2830.
- 12714 Pei, W., T. Ge, and B. Chang (2015). An effective neural network model for graph-based
12715 dependency parsing. In *Proceedings of the Association for Computational Linguistics (ACL)*,
12716 pp. 313–322.
- 12717 Peldszus, A. and M. Stede (2013). From argument diagrams to argumentation mining
12718 in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence
12719 (IJCINI)* 7(1), 1–31.
- 12720 Peldszus, A. and M. Stede (2015). An annotated corpus of argumentative microtexts. In
12721 *Proceedings of the First Conference on Argumentation*.
- 12722 Peng, F., F. Feng, and A. McCallum (2004). Chinese segmentation and new word detec-
12723 tion using conditional random fields. In *Proceedings of the International Conference on
12724 Computational Linguistics (COLING)*, pp. 562–568.
- 12725 Pennington, J., R. Socher, and C. Manning (2014). Glove: Global vectors for word repre-
12726 sentation. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*,
12727 pp. 1532–1543.

- 12728 Pereira, F. and Y. Schabes (1992). Inside-outside reestimation from partially bracketed
12729 corpora. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 128–
12730 135.
- 12731 Pereira, F. C. N. and S. M. Shieber (2002). *Prolog and natural-language analysis*. Microtome
12732 Publishing.
- 12733 Peters, M. E., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer
12734 (2018). Deep contextualized word representations. In *Proceedings of the North American
12735 Chapter of the Association for Computational Linguistics (NAACL)*.
- 12736 Peterson, W. W., T. G. Birdsall, and W. C. Fox (1954). The theory of signal detectability.
12737 *Transactions of the IRE professional group on information theory* 4(4), 171–212.
- 12738 Petrov, S., L. Barrett, R. Thibaux, and D. Klein (2006). Learning accurate, compact, and in-
12739 terpretable tree annotation. In *Proceedings of the Association for Computational Linguistics
12740 (ACL)*.
- 12741 Petrov, S., D. Das, and R. McDonald (2012). A universal part-of-speech tagset. In *Proceed-
12742 ings of the Language Resources and Evaluation Conference*.
- 12743 Petrov, S. and R. McDonald (2012). Overview of the 2012 shared task on parsing the web.
12744 In *Notes of the First Workshop on Syntactic Analysis of Non-Canonical Language (SANCL)*.
- 12745 Pinker, S. (2003). *The language instinct: How the mind creates language*. Penguin UK.
- 12746 Pinter, Y., R. Guthrie, and J. Eisenstein (2017). Mimicking word embeddings using
12747 subword RNNs. In *Proceedings of Empirical Methods for Natural Language Processing
12748 (EMNLP)*.
- 12749 Pitler, E., A. Louis, and A. Nenkova (2009). Automatic sense prediction for implicit dis-
12750 course relations in text. In *Proceedings of the Association for Computational Linguistics
12751 (ACL)*.
- 12752 Pitler, E. and A. Nenkova (2009). Using syntax to disambiguate explicit discourse con-
12753 nectives in text. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp.
12754 13–16.
- 12755 Pitler, E., M. Raghupathy, H. Mehta, A. Nenkova, A. Lee, and A. Joshi (2008). Easily iden-
12756 tifiable discourse relations. In *Proceedings of the International Conference on Computational
12757 Linguistics (COLING)*, pp. 87–90.
- 12758 Plank, B., A. Søgaard, and Y. Goldberg (2016). Multilingual part-of-speech tagging with
12759 bidirectional long short-term memory models and auxiliary loss. In *Proceedings of the
12760 Association for Computational Linguistics (ACL)*.

- 12761 Poesio, M., R. Stevenson, B. Di Eugenio, and J. Hitzeman (2004). Centering: A parametric
12762 theory and its instantiations. *Computational linguistics* 30(3), 309–363.
- 12763 Polanyi, L. and A. Zaenen (2006). Contextual valence shifters. In *Computing attitude and*
12764 *affect in text: Theory and applications*. Springer.
- 12765 Ponzetto, S. P. and M. Strube (2006). Exploiting semantic role labeling, wordnet and
12766 wikipedia for coreference resolution. In *Proceedings of the North American Chapter of*
12767 *the Association for Computational Linguistics (NAACL)*, pp. 192–199.
- 12768 Ponzetto, S. P. and M. Strube (2007). Knowledge derived from wikipedia for computing
12769 semantic relatedness. *Journal of Artificial Intelligence Research* 30, 181–212.
- 12770 Poon, H. and P. Domingos (2008). Joint unsupervised coreference resolution with markov
12771 logic. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp.
12772 650–659.
- 12773 Poon, H. and P. Domingos (2009). Unsupervised semantic parsing. In *Proceedings of Em-*
12774 *pirical Methods for Natural Language Processing (EMNLP)*, pp. 1–10.
- 12775 Popel, M., D. Marecek, J. Stepánek, D. Zeman, and Z. Zabokrtský (2013). Coordination
12776 structures in dependency treebanks. In *Proceedings of the Association for Computational*
12777 *Linguistics (ACL)*, pp. 517–527.
- 12778 Popescu, A.-M., O. Etzioni, and H. Kautz (2003). Towards a theory of natural language
12779 interfaces to databases. In *Proceedings of Intelligent User Interfaces (IUI)*, pp. 149–157.
- 12780 Poplack, S. (1980). Sometimes I'll start a sentence in Spanish y termino en Español: toward
12781 a typology of code-switching. *Linguistics* 18(7-8), 581–618.
- 12782 Porter, M. F. (1980). An algorithm for suffix stripping. *Program* 14(3), 130–137.
- 12783 Prabhakaran, V., O. Rambow, and M. Diab (2010). Automatic committed belief tagging.
12784 In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pp.
12785 1014–1022.
- 12786 Pradhan, S., X. Luo, M. Recasens, E. Hovy, V. Ng, and M. Strube (2014). Scoring corefer-
12787 ence partitions of predicted mentions: A reference implementation. In *Proceedings of the*
12788 *Association for Computational Linguistics (ACL)*, pp. 30–35.
- 12789 Pradhan, S., L. Ramshaw, M. Marcus, M. Palmer, R. Weischedel, and N. Xue (2011).
12790 CoNLL-2011 shared task: Modeling unrestricted coreference in OntoNotes. In *Proceed-
12791 ings of the Conference on Natural Language Learning (CoNLL)*, pp. 1–27.

- 12792 Pradhan, S., W. Ward, K. Hacioglu, J. H. Martin, and D. Jurafsky (2005). Semantic role
12793 labeling using different syntactic views. In *Proceedings of the Association for Computational
12794 Linguistics (ACL)*, pp. 581–588.
- 12795 Prasad, R., N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber (2008).
12796 The Penn Discourse Treebank 2.0. In *Proceedings of the Language Resources and Evaluation
12797 Conference*.
- 12798 Punyakanok, V., D. Roth, and W.-t. Yih (2008). The importance of syntactic parsing and
12799 inference in semantic role labeling. *Computational Linguistics* 34(2), 257–287.
- 12800 Pustejovsky, J., P. Hanks, R. Saurí, A. See, R. Gaizauskas, A. Setzer, D. Radov, B. Sundheim,
12801 D. Day, L. Ferro, et al. (2003). The timebank corpus. In *Proceedings of Corpus linguistics*,
12802 pp. 647–656.
- 12803 Pustejovsky, J., B. Ingria, R. Saurí, J. Castano, J. Littman, R. Gaizauskas, A. Setzer, G. Katz,
12804 and I. Mani (2005). The specification language TimeML. In *The language of time: A reader*,
12805 pp. 545–557. Oxford University Press.
- 12806 Qin, L., Z. Zhang, H. Zhao, Z. Hu, and E. Xing (2017). Adversarial connective-exploiting
12807 networks for implicit discourse relation classification. In *Proceedings of the Association
12808 for Computational Linguistics (ACL)*, pp. 1006–1017.
- 12809 Qiu, G., B. Liu, J. Bu, and C. Chen (2011). Opinion word expansion and target extraction
12810 through double propagation. *Computational linguistics* 37(1), 9–27.
- 12811 Quattoni, A., S. Wang, L.-P. Morency, M. Collins, and T. Darrell (2007). Hidden conditional
12812 random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29(10).
- 12813 Rahman, A. and V. Ng (2011). Narrowing the modeling gap: a cluster-ranking approach
12814 to coreference resolution. *Journal of Artificial Intelligence Research* 40, 469–521.
- 12815 Rajpurkar, P., J. Zhang, K. Lopyrev, and P. Liang (2016). Squad: 100,000+ questions for
12816 machine comprehension of text. In *Proceedings of Empirical Methods for Natural Language
12817 Processing (EMNLP)*, pp. 2383–2392.
- 12818 Ranzato, M., S. Chopra, M. Auli, and W. Zaremba (2016). Sequence level training with
12819 recurrent neural networks. In *Proceedings of the International Conference on Learning Rep-
12820 resentations (ICLR)*.
- 12821 Rao, D., D. Yarowsky, A. Shreevats, and M. Gupta (2010). Classifying latent user attributes
12822 in twitter. In *Proceedings of Workshop on Search and mining user-generated contents*.
- 12823 Ratinov, L. and D. Roth (2009). Design challenges and misconceptions in named entity
12824 recognition. In *Proceedings of the Conference on Natural Language Learning (CoNLL)*, pp.
12825 147–155.

- 12826 Ratinov, L., D. Roth, D. Downey, and M. Anderson (2011). Local and global algorithms
 12827 for disambiguation to wikipedia. In *Proceedings of the Association for Computational Lin-*
 12828 *guistics (ACL)*, pp. 1375–1384.
- 12829 Ratliff, N. D., J. A. Bagnell, and M. Zinkevich (2007). (approximate) subgradient methods
 12830 for structured prediction. In *Proceedings of Artificial Intelligence and Statistics (AISTATS)*,
 12831 pp. 380–387.
- 12832 Ratnaparkhi, A. (1996). A maximum entropy model for part-of-speech tagging. In *Pro-*
 12833 , pp. 133–142.
- 12834 Ratnaparkhi, A., J. Reynar, and S. Roukos (1994). A maximum entropy model for preposi-
 12835 *tional phrase attachment*. In *Proceedings of the workshop on Human Language Technology*,
 12836 pp. 250–255.
- 12837 Read, J. (2005). Using emoticons to reduce dependency in machine learning techniques for
 12838 sentiment classification. In *Proceedings of the ACL student research workshop*, pp. 43–48.
- 12839 Reisinger, D., R. Rudinger, F. Ferraro, C. Harman, K. Rawlins, and B. V. Durme (2015).
 12840 Semantic proto-roles. *Transactions of the Association for Computational Linguistics* 3, 475–
 12841 488.
- 12842 Reisinger, J. and R. J. Mooney (2010). Multi-prototype vector-space models of word mean-
 12843 *ing*. In *Proceedings of the North American Chapter of the Association for Computational Lin-*
 12844 *guistics (NAACL)*, pp. 109–117.
- 12845 Reiter, E. and R. Dale (2000). *Building natural language generation systems*. Cambridge
 12846 University Press.
- 12847 Resnik, P., M. B. Olsen, and M. Diab (1999). The bible as a parallel corpus: Annotating the
 12848 ‘Book of 2000 Tongues’. *Computers and the Humanities* 33(1-2), 129–153.
- 12849 Resnik, P. and N. A. Smith (2003). The web as a parallel corpus. *Computational Linguis-*
 12850 *tics* 29(3), 349–380.
- 12851 Ribeiro, F. N., M. Araújo, P. Gonçalves, M. A. Gonçalves, and F. Benevenuto (2016).
 12852 Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis meth-
 12853 ods. *EPJ Data Science* 5(1), 1–29.
- 12854 Richardson, M., C. J. Burges, and E. Renshaw (2013). MCTest: A challenge dataset for
 12855 the open-domain machine comprehension of text. In *Proceedings of Empirical Methods for*
 12856 *Natural Language Processing (EMNLP)*, pp. 193–203.
- 12857 Riedel, S., L. Yao, and A. McCallum (2010). Modeling relations and their mentions without
 12858 labeled text. In *Proceedings of the European Conference on Machine Learning and Principles*
 12859 *and Practice of Knowledge Discovery in Databases (ECML)*, pp. 148–163.

- 12860 Riedl, M. O. and R. M. Young (2010). Narrative planning: Balancing plot and character.
12861 *Journal of Artificial Intelligence Research* 39, 217–268.
- 12862 Rieser, V. and O. Lemon (2011). *Reinforcement learning for adaptive dialogue systems: a data-
12863 driven methodology for dialogue management and natural language generation*. Springer Sci-
12864 ence & Business Media.
- 12865 Riloff, E. (1996). Automatically generating extraction patterns from untagged text. In
12866 *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, pp. 1044–1049.
- 12867 Riloff, E. and J. Wiebe (2003). Learning extraction patterns for subjective expressions. In
12868 *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pp.
12869 105–112. Association for Computational Linguistics.
- 12870 Ritchie, G. (2001). Current directions in computational humour. *Artificial Intelligence Re-
12871 view* 16(2), 119–135.
- 12872 Ritter, A., C. Cherry, and W. B. Dolan (2011). Data-driven response generation in social
12873 media. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp.
12874 583–593.
- 12875 Ritter, A., S. Clark, Mausam, and O. Etzioni (2011). Named entity recognition in tweets: an
12876 experimental study. In *Proceedings of Empirical Methods for Natural Language Processing
(EMNLP)*.
- 12878 Roark, B., M. Saraclar, and M. Collins (2007). Discriminative n -gram language modeling.
12879 *Computer Speech & Language* 21(2), 373–392.
- 12880 Robert, C. and G. Casella (2013). *Monte Carlo statistical methods*. Springer Science & Busi-
12881 ness Media.
- 12882 Rosenfeld, R. (1996). A maximum entropy approach to adaptive statistical language mod-
12883 ellng. *Computer Speech & Language* 10(3), 187–228.
- 12884 Ross, S., G. Gordon, and D. Bagnell (2011). A reduction of imitation learning and struc-
12885 tured prediction to no-regret online learning. In *Proceedings of Artificial Intelligence and
12886 Statistics (AISTATS)*, pp. 627–635.
- 12887 Roy, N., J. Pineau, and S. Thrun (2000). Spoken dialogue management using probabilistic
12888 reasoning. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 93–
12889 100.
- 12890 Rudinger, R., J. Naradowsky, B. Leonard, and B. Van Durme (2018). Gender bias in coref-
12891 erence resolution. In *Proceedings of the North American Chapter of the Association for Com-
12892 putational Linguistics (NAACL)*.

- 12893 Rudnicky, A. and W. Xu (1999). An agenda-based dialog management architecture for
12894 spoken language systems. In *IEEE Automatic Speech Recognition and Understanding Work-*
12895 *shop*, Volume 13.
- 12896 Rush, A. M., S. Chopra, and J. Weston (2015). A neural attention model for abstractive sen-
12897 tence summarization. In *Proceedings of Empirical Methods for Natural Language Processing*
12898 (*EMNLP*), pp. 379–389.
- 12899 Rush, A. M., D. Sontag, M. Collins, and T. Jaakkola (2010). On dual decomposition and
12900 linear programming relaxations for natural language processing. In *Proceedings of Em-*
12901 *pirical Methods for Natural Language Processing (EMNLP)*, pp. 1–11.
- 12902 Russell, S. J. and P. Norvig (2009). *Artificial intelligence: a modern approach* (3rd ed.). Prentice
12903 Hall.
- 12904 Rutherford, A., V. Demberg, and N. Xue (2017). A systematic study of neural discourse
12905 models for implicit discourse relation. In *Proceedings of the European Chapter of the Asso-*
12906 *ciation for Computational Linguistics (EACL)*, pp. 281–291.
- 12907 Rutherford, A. T. and N. Xue (2014). Discovering implicit discourse relations through
12908 brown cluster pair representation and coreference patterns. In *Proceedings of the Euro-*
12909 *pean Chapter of the Association for Computational Linguistics (EACL)*.
- 12910 Sag, I. A., T. Baldwin, F. Bond, A. Copestake, and D. Flickinger (2002). Multiword expres-
12911 sions: A pain in the neck for nlp. In *International Conference on Intelligent Text Processing*
12912 *and Computational Linguistics*, pp. 1–15. Springer.
- 12913 Sagae, K. (2009). Analysis of discourse structure with syntactic dependencies and data-
12914 driven shift-reduce parsing. In *Proceedings of the 11th International Conference on Parsing*
12915 *Technologies*, pp. 81–84.
- 12916 Santos, C. D. and B. Zadrozny (2014). Learning character-level representations for part-of-
12917 speech tagging. In *Proceedings of the International Conference on Machine Learning (ICML)*,
12918 pp. 1818–1826.
- 12919 Sato, M.-A. and S. Ishii (2000). On-line em algorithm for the normalized gaussian network.
12920 *Neural computation* 12(2), 407–432.
- 12921 Saurí, R. and J. Pustejovsky (2009). Factbank: a corpus annotated with event factuality.
12922 *Language resources and evaluation* 43(3), 227.
- 12923 Saxe, A. M., J. L. McClelland, and S. Ganguli (2014). Exact solutions to the nonlinear
12924 dynamics of learning in deep linear neural networks. In *Proceedings of the International*
12925 *Conference on Learning Representations (ICLR)*.

- 12926 Schank, R. C. and R. Abelson (1977). *Scripts, goals, plans, and understanding*. Hillsdale, NJ:
12927 Erlbaum.
- 12928 Schapire, R. E. and Y. Singer (2000). BoosTexter: A boosting-based system for text categorization. *Machine learning* 39(2-3), 135–168.
- 12930 Schaul, T., S. Zhang, and Y. LeCun (2013). No more pesky learning rates. In *Proceedings of*
12931 *the International Conference on Machine Learning (ICML)*, pp. 343–351.
- 12932 Schnabel, T., I. Labutov, D. Mimno, and T. Joachims (2015). Evaluation methods for un-
12933 supervised word embeddings. In *Proceedings of Empirical Methods for Natural Language*
12934 *Processing (EMNLP)*, pp. 298–307.
- 12935 Schneider, N., J. Flanigan, and T. O’Gorman (2015). The logic of AMR: Practical, unified,
12936 graph-based sentence semantics for NLP. In *Proceedings of the North American Chapter of*
12937 *the Association for Computational Linguistics (NAACL)*, pp. 4–5.
- 12938 Schütze, H. (1998). Automatic word sense discrimination. *Computational linguistics* 24(1),
12939 97–123.
- 12940 Schwarm, S. E. and M. Ostendorf (2005). Reading level assessment using support vector
12941 machines and statistical language models. In *Proceedings of the Association for Compu-*
12942 *tational Linguistics (ACL)*, pp. 523–530.
- 12943 See, A., P. J. Liu, and C. D. Manning (2017). Get to the point: Summarization with pointer-
12944 generator networks. In *Proceedings of the Association for Computational Linguistics (ACL)*,
12945 pp. 1073–1083.
- 12946 Sennrich, R., B. Haddow, and A. Birch (2016). Neural machine translation of rare words
12947 with subword units. In *Proceedings of the Association for Computational Linguistics (ACL)*,
12948 pp. 1715–1725.
- 12949 Serban, I. V., A. Sordoni, Y. Bengio, A. C. Courville, and J. Pineau (2016). Building end-to-
12950 end dialogue systems using generative hierarchical neural network models. In *Proceed-*
12951 *ings of the National Conference on Artificial Intelligence (AAAI)*, pp. 3776–3784.
- 12952 Settles, B. (2012). Active learning. *Synthesis Lectures on Artificial Intelligence and Machine*
12953 *Learning* 6(1), 1–114.
- 12954 Shang, L., Z. Lu, and H. Li (2015). Neural responding machine for short-text conversation.
12955 In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 1577–1586.
- 12956 Shen, D. and M. Lapata (2007). Using semantic roles to improve question answering. In
12957 *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 12–21.

- 12958 Shen, S., Y. Cheng, Z. He, W. He, H. Wu, M. Sun, and Y. Liu (2016). Minimum risk training
 12959 for neural machine translation. In *Proceedings of the Association for Computational
 12960 Linguistics (ACL)*, pp. 1683–1692.
- 12961 Shen, W., J. Wang, and J. Han (2015). Entity linking with a knowledge base: Issues, tech-
 12962 niques, and solutions. *IEEE Transactions on Knowledge and Data Engineering* 27(2), 443–
 12963 460.
- 12964 Shieber, S. M. (1985). Evidence against the context-freeness of natural language. *Linguistics
 12965 and Philosophy* 8(3), 333–343.
- 12966 Siegelmann, H. T. and E. D. Sontag (1995). On the computational power of neural nets.
 12967 *Journal of computer and system sciences* 50(1), 132–150.
- 12968 Singh, S., A. Subramanya, F. Pereira, and A. McCallum (2011). Large-scale cross-
 12969 document coreference using distributed inference and hierarchical models. In *Proceed-
 12970 ings of the Association for Computational Linguistics (ACL)*, pp. 793–803.
- 12971 Sipser, M. (2012). *Introduction to the Theory of Computation*. Cengage Learning.
- 12972 Smith, D. A. and J. Eisner (2006). Minimum risk annealing for training log-linear models.
 12973 In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 787–794.
- 12974 Smith, D. A. and J. Eisner (2008). Dependency parsing by belief propagation. In *Proceed-
 12975 ings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 145–156.
- 12976 Smith, D. A. and N. A. Smith (2007). Probabilistic models of nonprojective dependency
 12977 trees. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp.
 12978 132–140.
- 12979 Smith, N. A. (2011). Linguistic structure prediction. *Synthesis Lectures on Human Language
 12980 Technologies* 4(2), 1–274.
- 12981 Snover, M., B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul (2006). A study of transla-
 12982 tion edit rate with targeted human annotation. In *Proceedings of Association for Machine
 12983 Translation in the Americas (AMTA)*.
- 12984 Snow, R., B. O’Connor, D. Jurafsky, and A. Y. Ng (2008). Cheap and fast—but is it good?:
 12985 evaluating non-expert annotations for natural language tasks. In *Proceedings of Empirical
 12986 Methods for Natural Language Processing (EMNLP)*, pp. 254–263.
- 12987 Snyder, B. and R. Barzilay (2007). Database-text alignment via structured multilabel classi-
 12988 fication. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*,
 12989 pp. 1713–1718.

- 12990 Socher, R., J. Bauer, C. D. Manning, and A. Y. Ng (2013). Parsing with compositional vector
12991 grammars. In *Proceedings of the Association for Computational Linguistics (ACL)*.
- 12992 Socher, R., B. Huval, C. D. Manning, and A. Y. Ng (2012). Semantic compositionality
12993 through recursive matrix-vector spaces. In *Proceedings of Empirical Methods for Natural
12994 Language Processing (EMNLP)*, pp. 1201–1211.
- 12995 Socher, R., A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts (2013).
12996 Recursive deep models for semantic compositionality over a sentiment treebank. In
12997 *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.
- 12998 Søgaard, A. (2013). Semi-supervised learning and domain adaptation in natural language
12999 processing. *Synthesis Lectures on Human Language Technologies* 6(2), 1–103.
- 13000 Solorio, T. and Y. Liu (2008). Learning to predict code-switching points. In *Proceedings of
13001 Empirical Methods for Natural Language Processing (EMNLP)*, pp. 973–981.
- 13002 Somasundaran, S., G. Namata, J. Wiebe, and L. Getoor (2009). Supervised and unsuper-
13003 vised methods in employing discourse relations for improving opinion polarity classi-
13004 fication. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.
- 13005 Somasundaran, S. and J. Wiebe (2009). Recognizing stances in online debates. In *Proceed-
13006 ings of the Association for Computational Linguistics (ACL)*, pp. 226–234.
- 13007 Song, L., B. Boots, S. M. Siddiqi, G. J. Gordon, and A. J. Smola (2010). Hilbert space
13008 embeddings of hidden markov models. In *Proceedings of the International Conference on
13009 Machine Learning (ICML)*, pp. 991–998.
- 13010 Song, L., Y. Zhang, X. Peng, Z. Wang, and D. Gildea (2016). AMR-to-text generation as
13011 a traveling salesman problem. In *Proceedings of Empirical Methods for Natural Language
13012 Processing (EMNLP)*, pp. 2084–2089.
- 13013 Soon, W. M., H. T. Ng, and D. C. Y. Lim (2001). A machine learning approach to corefer-
13014 ence resolution of noun phrases. *Computational linguistics* 27(4), 521–544.
- 13015 Sordoni, A., M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J.-Y. Nie, J. Gao, and B. Dolan
13016 (2015). A neural network approach to context-sensitive generation of conversational
13017 responses. In *Proceedings of the North American Chapter of the Association for Computational
13018 Linguistics (NAACL)*.
- 13019 Soriceut, R. and D. Marcu (2003). Sentence level discourse parsing using syntactic and
13020 lexical information. In *Proceedings of the North American Chapter of the Association for
13021 Computational Linguistics (NAACL)*, pp. 149–156.

- 13022 Sowa, J. F. (2000). *Knowledge representation: logical, philosophical, and computational foundations*. Pacific Grove, CA: Brooks/Cole.
- 13023
- 13024 Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application
13025 in retrieval. *Journal of documentation* 28(1), 11–21.
- 13026 Spitkovsky, V. I., H. Alshawi, D. Jurafsky, and C. D. Manning (2010). Viterbi training
13027 improves unsupervised dependency parsing. In *Proceedings of the Conference on Natural
13028 Language Learning (CoNLL)*, pp. 9–17.
- 13029 Sporleder, C. and M. Lapata (2005). Discourse chunking and its application to sen-
13030 tence compression. In *Proceedings of Empirical Methods for Natural Language Processing
13031 (EMNLP)*, pp. 257–264.
- 13032 Sproat, R., A. Black, S. Chen, S. Kumar, M. Ostendorf, and C. Richards (2001). Normaliza-
13033 tion of non-standard words. *Computer Speech & Language* 15(3), 287–333.
- 13034 Sproat, R., W. Gale, C. Shih, and N. Chang (1996). A stochastic finite-state word-
13035 segmentation algorithm for chinese. *Computational linguistics* 22(3), 377–404.
- 13036 Sra, S., S. Nowozin, and S. J. Wright (2012). *Optimization for machine learning*. MIT Press.
- 13037 Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov (2014).
13038 Dropout: A simple way to prevent neural networks from overfitting. *The Journal of
13039 Machine Learning Research* 15(1), 1929–1958.
- 13040 Srivastava, R. K., K. Greff, and J. Schmidhuber (2015). Training very deep networks. In
13041 *Neural Information Processing Systems (NIPS)*, pp. 2377–2385.
- 13042 Stab, C. and I. Gurevych (2014a). Annotating argument components and relations in per-
13043 suasive essays. In *Proceedings of the International Conference on Computational Linguistics
13044 (COLING)*, pp. 1501–1510.
- 13045 Stab, C. and I. Gurevych (2014b). Identifying argumentative discourse structures in per-
13046 suasive essays. In *Proceedings of Empirical Methods for Natural Language Processing
13047 (EMNLP)*, pp. 46–56.
- 13048 Stede, M. (2011, nov). *Discourse Processing*, Volume 4 of *Synthesis Lectures on Human Lan-
13049 guage Technologies*. Morgan & Claypool Publishers.
- 13050 Steedman, M. and J. Baldridge (2011). Combinatory categorial grammar. In *Non-
13051 Transformational Syntax: Formal and Explicit Models of Grammar*. Wiley-Blackwell.
- 13052 Stenetorp, P., S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii (2012). Brat: a web-
13053 based tool for nlp-assisted text annotation. In *Proceedings of the European Chapter of the
13054 Association for Computational Linguistics (EACL)*, pp. 102–107.

- 13055 Stern, M., J. Andreas, and D. Klein (2017). A minimal span-based neural constituency
13056 parser. In *Proceedings of the Association for Computational Linguistics (ACL)*.
- 13057 Stolcke, A., K. Ries, N. Coccato, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin,
13058 C. Van Ess-Dykema, and M. Meteer (2000). Dialogue act modeling for automatic tag-
13059 ging and recognition of conversational speech. *Computational linguistics* 26(3), 339–373.
- 13060 Stone, P. J. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. The MIT
13061 Press.
- 13062 Stoyanov, V., N. Gilbert, C. Cardie, and E. Riloff (2009). Conundrums in noun phrase
13063 coreference resolution: Making sense of the state-of-the-art. In *Proceedings of the Associa-*
13064 *tion for Computational Linguistics (ACL)*, pp. 656–664.
- 13065 Strang, G. (2016). *Introduction to linear algebra* (Fifth ed.). Wellesley, MA: Wellesley-
13066 Cambridge Press.
- 13067 Strubell, E., P. Verga, D. Belanger, and A. McCallum (2017). Fast and accurate entity recog-
13068 nition with iterated dilated convolutions. In *Proceedings of Empirical Methods for Natural*
13069 *Language Processing (EMNLP)*.
- 13070 Suchanek, F. M., G. Kasneci, and G. Weikum (2007). Yago: a core of semantic knowledge.
13071 In *Proceedings of the Conference on World-Wide Web (WWW)*, pp. 697–706.
- 13072 Sun, X., T. Matsuzaki, D. Okanohara, and J. Tsujii (2009). Latent variable perceptron algo-
13073 rithm for structured classification. In *Proceedings of the International Joint Conference on*
13074 *Artificial Intelligence (IJCAI)*, pp. 1236–1242.
- 13075 Sun, Y., L. Lin, D. Tang, N. Yang, Z. Ji, and X. Wang (2015). Modeling mention, context and
13076 entity with neural networks for entity disambiguation. In *Proceedings of the International*
13077 *Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1333–1339.
- 13078 Sundermeyer, M., R. Schlüter, and H. Ney (2012). LSTM neural networks for language
13079 modeling. In *Proceedings of the International Speech Communication Association (INTER-*
13080 *SPEECH)*.
- 13081 Surdeanu, M., J. Tibshirani, R. Nallapati, and C. D. Manning (2012). Multi-instance multi-
13082 label learning for relation extraction. In *Proceedings of Empirical Methods for Natural Lan-*
13083 *guage Processing (EMNLP)*, pp. 455–465.
- 13084 Sutskever, I., O. Vinyals, and Q. V. Le (2014). Sequence to sequence learning with neural
13085 networks. In *Neural Information Processing Systems (NIPS)*, pp. 3104–3112.
- 13086 Sutton, R. S. and A. G. Barto (1998). *Reinforcement learning: An introduction*, Volume 1. MIT
13087 press Cambridge.

- 13088 Sutton, R. S., D. A. McAllester, S. P. Singh, and Y. Mansour (2000). Policy gradient methods
 13089 for reinforcement learning with function approximation. In *Neural Information Processing-
 13090 ing Systems (NIPS)*, pp. 1057–1063.
- 13091 Suzuki, J., S. Takase, H. Kamigaito, M. Morishita, and M. Nagata (2018). An empirical
 13092 study of building a strong baseline for constituency parsing. In *Proceedings of the Asso-
 13093 ciation for Computational Linguistics (ACL)*, pp. 612–618.
- 13094 Sweeney, L. (2013). Discrimination in online ad delivery. *Queue* 11(3), 10.
- 13095 Taboada, M., J. Brooke, M. Tofiloski, K. Voll, and M. Stede (2011). Lexicon-based methods
 13096 for sentiment analysis. *Computational linguistics* 37(2), 267–307.
- 13097 Taboada, M. and W. C. Mann (2006). Rhetorical structure theory: Looking back and mov-
 13098 ing ahead. *Discourse studies* 8(3), 423–459.
- 13099 Täckström, O., K. Ganchev, and D. Das (2015). Efficient inference and structured learning
 13100 for semantic role labeling. *Transactions of the Association for Computational Linguistics* 3,
 13101 29–41.
- 13102 Täckström, O., R. McDonald, and J. Uszkoreit (2012). Cross-lingual word clusters for
 13103 direct transfer of linguistic structure. In *Proceedings of the North American Chapter of the
 13104 Association for Computational Linguistics (NAACL)*, pp. 477–487.
- 13105 Tang, D., B. Qin, and T. Liu (2015). Document modeling with gated recurrent neural net-
 13106 work for sentiment classification. In *Proceedings of Empirical Methods for Natural Language
 13107 Processing (EMNLP)*, pp. 1422–1432.
- 13108 Taskar, B., C. Guestrin, and D. Koller (2003). Max-margin markov networks. In *Neural
 13109 Information Processing Systems (NIPS)*.
- 13110 Tausczik, Y. R. and J. W. Pennebaker (2010). The psychological meaning of words: LIWC
 13111 and computerized text analysis methods. *Journal of Language and Social Psychology* 29(1),
 13112 24–54.
- 13113 Teh, Y. W. (2006). A hierarchical bayesian language model based on pitman-yor processes.
 13114 In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 985–992.
- 13115 Tesnière, L. (1966). *Éléments de syntaxe structurale* (second ed.). Paris: Klincksieck.
- 13116 Teufel, S., J. Carletta, and M. Moens (1999). An annotation scheme for discourse-level
 13117 argumentation in research articles. In *Proceedings of the European Chapter of the Association
 13118 for Computational Linguistics (EACL)*, pp. 110–117.
- 13119 Teufel, S. and M. Moens (2002). Summarizing scientific articles: experiments with rele-
 13120 vance and rhetorical status. *Computational linguistics* 28(4), 409–445.

- 13121 Thomas, M., B. Pang, and L. Lee (2006). Get out the vote: Determining support or opposition
13122 from Congressional floor-debate transcripts. In *Proceedings of Empirical Methods for*
13123 *Natural Language Processing (EMNLP)*, pp. 327–335.
- 13124 Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal*
13125 *Statistical Society. Series B (Methodological)*, 267–288.
- 13126 Titov, I. and J. Henderson (2007). Constituent parsing with incremental sigmoid belief
13127 networks. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 632–
13128 639.
- 13129 Toutanova, K., D. Klein, C. D. Manning, and Y. Singer (2003). Feature-rich part-of-speech
13130 tagging with a cyclic dependency network. In *Proceedings of the North American Chapter*
13131 *of the Association for Computational Linguistics (NAACL)*.
- 13132 Trivedi, R. and J. Eisenstein (2013). Discourse connectors for latent subjectivity in senti-
13133 ment analysis. In *Proceedings of the North American Chapter of the Association for Compu-*
13134 *tational Linguistics (NAACL)*, pp. 808–813.
- 13135 Tromble, R. W. and J. Eisner (2006). A fast finite-state relaxation method for enforcing
13136 global constraints on sequence decoding. In *Proceedings of the North American Chapter of*
13137 *the Association for Computational Linguistics (NAACL)*.
- 13138 Tschantaridis, I., T. Hofmann, T. Joachims, and Y. Altun (2004). Support vector machine
13139 learning for interdependent and structured output spaces. In *Proceedings of the Interna-*
13140 *tional Conference on Machine Learning (ICML)*.
- 13141 Tsvetkov, Y., M. Faruqui, W. Ling, G. Lample, and C. Dyer (2015). Evaluation of word
13142 vector representations by subspace alignment. In *Proceedings of Empirical Methods for*
13143 *Natural Language Processing (EMNLP)*, pp. 2049–2054.
- 13144 Tu, Z., Z. Lu, Y. Liu, X. Liu, and H. Li (2016). Modeling coverage for neural machine
13145 translation. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 76–
13146 85.
- 13147 Turian, J., L. Ratinov, and Y. Bengio (2010). Word representations: a simple and general
13148 method for semi-supervised learning. In *Proceedings of the Association for Computational*
13149 *Linguistics (ACL)*, pp. 384–394.
- 13150 Turing, A. M. (2009). Computing machinery and intelligence. In R. Epstein, G. Roberts,
13151 and G. Beber (Eds.), *Parsing the Turing Test*, pp. 23–65. Springer.
- 13152 Turney, P. D. and P. Pantel (2010). From frequency to meaning: Vector space models of
13153 semantics. *Journal of Artificial Intelligence Research* 37, 141–188.

- 13154 Tutin, A. and R. Kittredge (1992). Lexical choice in context: generating procedural texts.
 13155 In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pp.
 13156 763–769.
- 13157 Twain, M. (1997). *A Tramp Abroad*. New York: Penguin.
- 13158 Tzeng, E., J. Hoffman, T. Darrell, and K. Saenko (2015). Simultaneous deep transfer across
 13159 domains and tasks. In *Proceedings of the International Conference on Computer Vision*
 13160 (*ICCV*), pp. 4068–4076.
- 13161 Usunier, N., D. Buffoni, and P. Gallinari (2009). Ranking with ordered weighted pairwise
 13162 classification. In *Proceedings of the International Conference on Machine Learning (ICML)*,
 13163 pp. 1057–1064.
- 13164 Uthus, D. C. and D. W. Aha (2013). The ubuntu chat corpus for multiparicipant chat
 13165 analysis. In *AAAI Spring Symposium: Analyzing Microtext*, Volume 13, pp. 01.
- 13166 Utiyama, M. and H. Isahara (2001). A statistical model for domain-independent text seg-
 13167 mentation. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 499–
 13168 506.
- 13169 Utiyama, M. and H. Isahara (2007). A comparison of pivot methods for phrase-based sta-
 13170 tistical machine translation. In *Proceedings of the North American Chapter of the Association*
 13171 *for Computational Linguistics (NAACL)*, pp. 484–491.
- 13172 Uzuner, Ö., X. Zhang, and T. Sibanda (2009). Machine learning and rule-based approaches
 13173 to assertion classification. *Journal of the American Medical Informatics Association* 16(1),
 13174 109–115.
- 13175 Vadas, D. and J. R. Curran (2011). Parsing noun phrases in the penn treebank. *Compu-
 13176 tational Linguistics* 37(4), 753–809.
- 13177 Van Eynde, F. (2006). NP-internal agreement and the structure of the noun phrase. *Journal*
 13178 *of Linguistics* 42(1), 139–186.
- 13179 Van Gael, J., A. Vlachos, and Z. Ghahramani (2009). The infinite hmm for unsuper-
 13180 vised pos tagging. In *Proceedings of Empirical Methods for Natural Language Processing*
 13181 (*EMNLP*), pp. 678–687.
- 13182 Vaswani, A., S. Bengio, E. Brevdo, F. Chollet, A. N. Gomez, S. Gouws, L. Jones, L. Kaiser,
 13183 N. Kalchbrenner, N. Parmar, R. Sepassi, N. Shazeer, and J. Uszkoreit (2018). Ten-
 13184 sor2tensor for neural machine translation. *CoRR abs/1803.07416*.
- 13185 Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and
 13186 I. Polosukhin (2017). Attention is all you need. In *Neural Information Processing Systems*
 13187 (*NIPS*), pp. 6000–6010.

- 13188 Velldal, E., L. Øvrelid, J. Read, and S. Oepen (2012). Speculation and negation: Rules,
13189 rankers, and the role of syntax. *Computational linguistics* 38(2), 369–410.
- 13190 Versley, Y. (2011). Towards finer-grained tagging of discourse connectives. In *Proceedings*
13191 of the Workshop Beyond Semantics: Corpus-based Investigations of Pragmatic and Discourse
13192 Phenomena, pp. 2–63.
- 13193 Vilain, M., J. Burger, J. Aberdeen, D. Connolly, and L. Hirschman (1995). A model-
13194 theoretic coreference scoring scheme. In *Proceedings of the 6th conference on Message*
13195 *understanding*, pp. 45–52. Association for Computational Linguistics.
- 13196 Vincent, P., H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol (2010). Stacked de-
13197 noising autoencoders: Learning useful representations in a deep network with a local
13198 denoising criterion. *The Journal of Machine Learning Research* 11, 3371–3408.
- 13199 Vincze, V., G. Szarvas, R. Farkas, G. Móra, and J. Csirik (2008). The bioscope corpus:
13200 biomedical texts annotated for uncertainty, negation and their scopes. *BMC bioinformat-
13201 ics* 9(11), S9.
- 13202 Vinyals, O., A. Toshev, S. Bengio, and D. Erhan (2015). Show and tell: A neural image cap-
13203 tion generator. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference*
13204 on, pp. 3156–3164. IEEE.
- 13205 Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum
13206 decoding algorithm. *IEEE transactions on Information Theory* 13(2), 260–269.
- 13207 Voll, K. and M. Taboada (2007). Not all words are created equal: Extracting semantic
13208 orientation as a function of adjective relevance. In *Proceedings of Australian Conference*
13209 on Artificial Intelligence.
- 13210 Wager, S., S. Wang, and P. S. Liang (2013). Dropout training as adaptive regularization. In
13211 *Neural Information Processing Systems (NIPS)*, pp. 351–359.
- 13212 Wainwright, M. J. and M. I. Jordan (2008). Graphical models, exponential families, and
13213 variational inference. *Foundations and Trends® in Machine Learning* 1(1-2), 1–305.
- 13214 Walker, M. A. (2000). An application of reinforcement learning to dialogue strategy selec-
13215 tion in a spoken dialogue system for email. *Journal of Artificial Intelligence Research* 12,
13216 387–416.
- 13217 Walker, M. A., J. E. Cahn, and S. J. Whittaker (1997). Improvising linguistic style: Social
13218 and affective bases for agent personality. In *Proceedings of the first international conference*
13219 on Autonomous agents

- 13220 Wang, C., N. Xue, and S. Pradhan (2015). A Transition-based Algorithm for AMR Parsing.
 13221 In *Proceedings of the North American Chapter of the Association for Computational Linguistics
 (NAACL)*, pp. 366–375.
- 13223 Wang, H., T. Onishi, K. Gimpel, and D. McAllester (2017). Emergent predication structure
 13224 in hidden state vectors of neural readers. In *Proceedings of the 2nd Workshop on Repre-
 13225 sentation Learning for NLP*, pp. 26–36.
- 13226 Weaver, W. (1955). Translation. *Machine translation of languages* 14, 15–23.
- 13227 Webber, B. (2004). D-LTAG: extending lexicalized TAG to discourse. *Cognitive Sci-
 13228 ence* 28(5), 751–779.
- 13229 Webber, B., M. Egg, and V. Kordoni (2012). Discourse structure and language technology.
 13230 *Journal of Natural Language Engineering* 18(4), 437–490.
- 13231 Webber, B. and A. Joshi (2012). Discourse structure and computation: past, present and
 13232 future. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Dis-
 13233 coveries*, pp. 42–54. Association for Computational Linguistics.
- 13234 Wei, G. C. and M. A. Tanner (1990). A Monte Carlo implementation of the EM algorithm
 13235 and the poor man’s data augmentation algorithms. *Journal of the American Statistical
 13236 Association* 85(411), 699–704.
- 13237 Weinberger, K., A. Dasgupta, J. Langford, A. Smola, and J. Attenberg (2009). Feature
 13238 hashing for large scale multitask learning. In *Proceedings of the International Conference
 13239 on Machine Learning (ICML)*, pp. 1113–1120.
- 13240 Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language
 13241 communication between man and machine. *Communications of the ACM* 9(1), 36–45.
- 13242 Wellner, B. and J. Pustejovsky (2007). Automatically identifying the arguments of dis-
 13243 course connectives. In *Proceedings of Empirical Methods for Natural Language Processing
 13244 (EMNLP)*, pp. 92–101.
- 13245 Wen, T.-H., M. Gasic, N. Mrkšić, P.-H. Su, D. Vandyke, and S. Young (2015). Semantically
 13246 conditioned lstm-based natural language generation for spoken dialogue systems. In
 13247 *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 1711–1721.
- 13248 Weston, J., S. Bengio, and N. Usunier (2011). Wsabie: Scaling up to large vocabulary im-
 13249 age annotation. In *Proceedings of the International Joint Conference on Artificial Intelligence
 13250 (IJCAI)*, pp. 2764–2770.
- 13251 Wiebe, J., T. Wilson, and C. Cardie (2005). Annotating expressions of opinions and emo-
 13252 tions in language. *Language resources and evaluation* 39(2), 165–210.

- 13253 Wieting, J., M. Bansal, K. Gimpel, and K. Livescu (2016a). CHARAGRAM: Embedding
13254 words and sentences via character n-grams. In *Proceedings of Empirical Methods for Nat-*
13255 *ural Language Processing (EMNLP)*, pp. 1504–1515.
- 13256 Wieting, J., M. Bansal, K. Gimpel, and K. Livescu (2016b). Towards universal paraphrastic
13257 sentence embeddings. In *Proceedings of the International Conference on Learning Represen-*
13258 *tations (ICLR)*.
- 13259 Williams, J. D. and S. Young (2007). Partially observable markov decision processes for
13260 spoken dialog systems. *Computer Speech & Language* 21(2), 393–422.
- 13261 Williams, P., R. Sennrich, M. Post, and P. Koehn (2016). Syntax-based statistical machine
13262 translation. *Synthesis Lectures on Human Language Technologies* 9(4), 1–208.
- 13263 Wilson, T., J. Wiebe, and P. Hoffmann (2005). Recognizing contextual polarity in phrase-
13264 level sentiment analysis. In *Proceedings of Empirical Methods for Natural Language Pro-*
13265 *cessing (EMNLP)*, pp. 347–354.
- 13266 Winograd, T. (1972). Understanding natural language. *Cognitive psychology* 3(1), 1–191.
- 13267 Wiseman, S., A. M. Rush, and S. M. Shieber (2016). Learning global features for corefer-
13268 ence resolution. In *Proceedings of the North American Chapter of the Association for Compu-*
13269 *tational Linguistics (NAACL)*, pp. 994–1004.
- 13270 Wiseman, S., S. Shieber, and A. Rush (2017). Challenges in data-to-document generation.
13271 In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 2253–
13272 2263.
- 13273 Wiseman, S. J., A. M. Rush, S. M. Shieber, and J. Weston (2015). Learning anaphoricity and
13274 antecedent ranking features for coreference resolution. In *Proceedings of the Association*
13275 *for Computational Linguistics (ACL)*.
- 13276 Wolf, F. and E. Gibson (2005). Representing discourse coherence: A corpus-based study.
13277 *Computational Linguistics* 31(2), 249–287.
- 13278 Wolfe, T., M. Dredze, and B. Van Durme (2017). Pocket knowledge base population. In
13279 *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 305–310.
- 13280 Wong, Y. W. and R. Mooney (2007). Generation by inverting a semantic parser that uses
13281 statistical machine translation. In *Proceedings of the North American Chapter of the Associa-*
13282 *tion for Computational Linguistics (NAACL)*, pp. 172–179.
- 13283 Wong, Y. W. and R. J. Mooney (2006). Learning for semantic parsing with statistical ma-
13284 chine translation. In *Proceedings of the North American Chapter of the Association for Com-*
13285 *putational Linguistics (NAACL)*, pp. 439–446.

- 13286 Wu, B. Y. and K.-M. Chao (2004). *Spanning trees and optimization problems*. CRC Press.
- 13287 Wu, D. (1997). Stochastic inversion transduction grammars and bilingual parsing of par-
13288 allel corpora. *Computational linguistics* 23(3), 377–403.
- 13289 Wu, F. and D. S. Weld (2010). Open information extraction using wikipedia. In *Proceedings*
13290 of the Association for Computational Linguistics (ACL), pp. 118–127.
- 13291 Wu, X., R. Ward, and L. Bottou (2018). WNGrad: Learn the learning rate in gradient
13292 descent. *arXiv preprint arXiv:1803.02865*.
- 13293 Wu, Y., M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao,
13294 Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Łukasz Kaiser, S. Gouws,
13295 Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young,
13296 J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean (2016).
13297 Google’s neural machine translation system: Bridging the gap between human and ma-
13298 chine translation. *CoRR abs/1609.08144*.
- 13299 Xia, F. (2000). The part-of-speech tagging guidelines for the Penn Chinese Treebank (3.0).
13300 Technical report, University of Pennsylvania Institute for Research in Cognitive Science.
- 13301 Xu, K., J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio
13302 (2015). Show, attend and tell: Neural image caption generation with visual attention.
13303 In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 2048–2057.
- 13304 Xu, W., X. Liu, and Y. Gong (2003). Document clustering based on non-negative matrix
13305 factorization. In *Proceedings of ACM SIGIR conference on Research and development in in-*
13306 *formation retrieval*, pp. 267–273.
- 13307 Xu, Y., L. Mou, G. Li, Y. Chen, H. Peng, and Z. Jin (2015). Classifying relations via long
13308 short term memory networks along shortest dependency paths. In *Proceedings of Empir-*
13309 *ical Methods for Natural Language Processing (EMNLP)*, pp. 1785–1794.
- 13310 Xuan Bach, N., N. L. Minh, and A. Shimazu (2012). A reranking model for discourse seg-
13311 mentation using subtree features. In *Proceedings of the Special Interest Group on Discourse*
13312 *and Dialogue (SIGDIAL)*.
- 13313 Xue, N. et al. (2003). Chinese word segmentation as character tagging. *Computational*
13314 *Linguistics and Chinese Language Processing* 8(1), 29–48.
- 13315 Xue, N., H. T. Ng, S. Pradhan, R. Prasad, C. Bryant, and A. T. Rutherford (2015). The
13316 CoNLL-2015 shared task on shallow discourse parsing. In *Proceedings of the Conference*
13317 *on Natural Language Learning (CoNLL)*.

- 13318 Yamada, H. and Y. Matsumoto (2003). Statistical dependency analysis with support vector
13319 machines. In *Proceedings of the 8th International Workshop on Parsing Technologies (IWPT)*,
13320 pp. 195–206.
- 13321 Yamada, K. and K. Knight (2001). A syntax-based statistical translation model. In *Proceed-
13322 ings of the Association for Computational Linguistics (ACL)*, pp. 523–530.
- 13323 Yang, B. and C. Cardie (2014). Context-aware learning for sentence-level sentiment anal-
13324 ysis with posterior regularization. In *Proceedings of the Association for Computational Lin-
13325 guistics (ACL)*.
- 13326 Yang, Y., M.-W. Chang, and J. Eisenstein (2016). Toward socially-infused information ex-
13327 traction: Embedding authors, mentions, and entities. In *Proceedings of Empirical Methods
13328 for Natural Language Processing (EMNLP)*.
- 13329 Yang, Y. and J. Eisenstein (2013). A log-linear model for unsupervised text normalization.
13330 In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.
- 13331 Yang, Y. and J. Eisenstein (2015). Unsupervised multi-domain adaptation with feature em-
13332 beddings. In *Proceedings of the North American Chapter of the Association for Computational
13333 Linguistics (NAACL)*.
- 13334 Yang, Y., W.-t. Yih, and C. Meek (2015). WikiQA: A challenge dataset for open-domain
13335 question answering. In *Proceedings of Empirical Methods for Natural Language Processing
13336 (EMNLP)*, pp. 2013–2018.
- 13337 Yannakoudakis, H., T. Briscoe, and B. Medlock (2011). A new dataset and method for
13338 automatically grading esol texts. In *Proceedings of the Association for Computational Lin-
13339 guistics (ACL)*, pp. 180–189.
- 13340 Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised meth-
13341 ods. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 189–196.
13342 Association for Computational Linguistics.
- 13343 Yee, L. C. and T. Y. Jones (2012, March). Apple CEO in China mission to clear up problems.
13344 *Reuters*. retrieved on March 26, 2017.
- 13345 Yi, Y., C.-Y. Lai, S. Petrov, and K. Keutzer (2011). Efficient parallel cky parsing on gpus. In
13346 *Proceedings of the 12th International Conference on Parsing Technologies*, pp. 175–185.
- 13347 Yu, C.-N. J. and T. Joachims (2009). Learning structural svms with latent variables. In
13348 *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 1169–1176.
- 13349 Yu, F. and V. Koltun (2016). Multi-scale context aggregation by dilated convolutions. In
13350 *Proceedings of the International Conference on Learning Representations (ICLR)*.

- 13351 Zaidan, O. F. and C. Callison-Burch (2011). Crowdsourcing translation: Professional qual-
 13352 ity from non-professionals. In *Proceedings of the Association for Computational Linguistics*
 13353 (ACL), pp. 1220–1229.
- 13354 Zaremba, W., I. Sutskever, and O. Vinyals. Recurrent neural network regularization. *arXiv*
 13355 preprint arXiv:1409.2329.
- 13356 Zeiler, M. D. (2012). ADADELTA: an adaptive learning rate method. *arXiv preprint*
 13357 arXiv:1212.5701.
- 13358 Zelenko, D., C. Aone, and A. Richardella (2003). Kernel methods for relation extraction.
 13359 *The Journal of Machine Learning Research* 3, 1083–1106.
- 13360 Zelle, J. M. and R. J. Mooney (1996). Learning to parse database queries using induc-
 13361 tive logic programming. In *Proceedings of the National Conference on Artificial Intelligence*
 13362 (AAAI), pp. 1050–1055.
- 13363 Zeng, D., K. Liu, S. Lai, G. Zhou, and J. Zhao (2014). Relation classification via convolu-
 13364 tional deep neural network. In *Proceedings of the International Conference on Computational*
 13365 *Linguistics* (COLING), pp. 2335–2344.
- 13366 Zettlemoyer, L. S. and M. Collins (2005). Learning to map sentences to logical form: Struc-
 13367 tured classification with probabilistic categorial grammars. In *Proceedings of Uncertainty*
 13368 in Artificial Intelligence (UAI).
- 13369 Zhang, C., S. Bengio, M. Hardt, B. Recht, and O. Vinyals (2017). Understanding deep
 13370 learning requires rethinking generalization. In *Proceedings of the International Conference*
 13371 on Learning Representations (ICLR).
- 13372 Zhang, X., J. Zhao, and Y. LeCun (2015). Character-level convolutional networks for text
 13373 classification. In *Neural Information Processing Systems* (NIPS), pp. 649–657.
- 13374 Zhang, Y. and S. Clark (2008). A tale of two parsers: investigating and combining graph-
 13375 based and transition-based dependency parsing using beam-search. In *Proceedings of*
 13376 *Empirical Methods for Natural Language Processing* (EMNLP), pp. 562–571.
- 13377 Zhang, Y., T. Lei, R. Barzilay, T. Jaakkola, and A. Globerson (2014). Steps to excellence:
 13378 Simple inference with refined scoring of dependency trees. In *Proceedings of the Associa-*
 13379 *tion for Computational Linguistics* (ACL), pp. 197–207.
- 13380 Zhang, Y. and J. Nivre (2011). Transition-based dependency parsing with rich non-local
 13381 features. In *Proceedings of the Association for Computational Linguistics* (ACL), pp. 188–193.
- 13382 Zhang, Z. (2017). A note on counting dependency trees. *arXiv preprint* arXiv:1708.08789.

- 13383 Zhao, J., T. Wang, M. Yatskar, e. V. Ordonez, and K.-W. Chang (2018). Gender bias in
13384 coreference resolution: Evaluation and debiasing methods. In *Proceedings of the North*
13385 *American Chapter of the Association for Computational Linguistics (NAACL)*.
- 13386 Zhou, J. and W. Xu (2015). End-to-end learning of semantic role labeling using recurrent
13387 neural networks. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp.
13388 1127–1137.
- 13389 Zhu, J., Z. Nie, X. Liu, B. Zhang, and J.-R. Wen (2009). Statsnowball: a statistical approach
13390 to extracting entity relationships. In *Proceedings of the Conference on World-Wide Web*
13391 (*WWW*), pp. 101–110.
- 13392 Zhu, X., Z. Ghahramani, and J. D. Lafferty (2003). Semi-supervised learning using gaussian
13393 fields and harmonic functions. In *Proceedings of the International Conference on Ma-*
13394 *chine Learning (ICML)*, pp. 912–919.
- 13395 Zhu, X. and A. B. Goldberg (2009). Introduction to semi-supervised learning. *Synthesis*
13396 *lectures on artificial intelligence and machine learning* 3(1), 1–130.
- 13397 Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Addison-Wesley.
- 13398 Zirn, C., M. Niepert, H. Stuckenschmidt, and M. Strube (2011). Fine-grained sentiment
13399 analysis with structural features. In *Proceedings of the International Joint Conference on*
13400 *Artificial Intelligence (IJCAI)*, pp. 336–344.
- 13401 Zou, W. Y., R. Socher, D. Cer, and C. D. Manning (2013). Bilingual word embeddings
13402 for phrase-based machine translation. In *Proceedings of Empirical Methods for Natural*
13403 *Language Processing (EMNLP)*, pp. 1393–1398.

13404 Index

- 13405 α -conversion, 295
13406 β -reduction, β -conversion, 293
13407 *n*-gram, 24
13408 language model, 198
13409 *p*-value, 84
13410 one-tailed, 85
13411 *F*-MEASURE, 82
13412 balanced, 83
13413 macro, 82
13414 micro, 83
13415 WORDNET, 9, 74
13416 BLEU, 433
13417 METEOR, 435
13418 RIBES, 435
13419 semantics
13420 extra-propositional, 422
13421 ablation test, 84
13422 Abstract Meaning Representation
13423 (AMR), 307, 319
13424 accuracy, 23, 81
13425 action, in reinforcement learning, 365
13426 AdaGrad, 40, 61
13427 adequacy, in translation, 433
13428 adjectives, 177
13429 adjuncts, 306
13430 adpositions, 178
13431 adverbs, 177
13432 affix, 195
13433 inflectional, 78
13434 agent (thematic role), 308
13435 alignment
13436 in machine translation, 432, 437
13437 in semantic parsing, 321
13438 in text generation, 459
13439 Amazon Mechanical Turk, 91
13440 ambiguity, 210, 218
13441 attachment, 257
13442 derivational, 222
13443 spurious, 222, 269, 298
13444 syntactic, 229
13445 anaphoricity, 360
13446 anchored productions, 232
13447 animacy, 307
13448 annealing, 452
13449 antecedent mention, 351, 359
13450 antonymy, 74, 281
13451 apophony, 194
13452 area under the curve (AUC), 83
13453 argumentation, 392
13454 arguments, 403
13455 article, 182
13456 aspect, 177
13457 aspect-based opinion mining, 71
13458 autoencoder, 345
13459 denoising, 345, 408
13460 variational, 465
13461 automated theorem provers, 288
13462 automatic differentiation, 56
13463 auxiliary verbs, 178
13464 average mutual information, 333
13465 averaged perceptron, 27

- 13466 backchannel, 187
 13467 backoff, 130
 Katz, 130
 13469 backpropagation, 55
 through time, 136
 13471 backward recurrence, 165, 166
 13472 backward-looking center, 382
 13473 bag of words, 13
 13474 balanced test set, 81
 13475 batch normalization, 61
 13476 Baum-Welch algorithm, 170
 13477 Bayes' rule, 478
 13478 Bayesian nonparametrics, 103, 249
 13479 beam sampling, 172
 13480 beam search, 272
 in coreference resolution, 363
 in machine translation, 450, 451
 13483 Bell number, 362
 13484 bias, 23
 13485 bias-variance tradeoff, 23, 127, 129
 13486 bigrams, 24, 70
 13487 bilinear product, 330
 13488 binarization, 211, 228
 13489 binomial
 distribution, 85
 random variable, 482
 13492 test, 84
 13493 BIO notation, 184, 317
 13494 biomedical natural language processing, 183
 13495 Brown clusters, 327, 332
 13496 byte-pair encoding, 343, 449
 13497 Bonferroni correction, 87
 13498 boolean semiring, 199
 13499 boosting, 48
 13500 bootstrap samples, 86
 13501 brevity penalty, 434
 13502 Catalan number, 225
 13506 cataphora, 352
 13507 center embedding, 207
 13508 centering theory, 355, 382
 13509 character-level language models, 141
 13510 chatbots, 470
 13511 Chomsky Normal Form (CNF), 211
 13512 Chu-Liu-Edmonds algorithm, 264
 13513 CKY algorithm, 226
 13514 class imbalance, 81
 13515 classification, 13
 large margin, 30
 lexicon-based, 72
 weights, 14
 13519 closed-vocabulary, 141
 13520 closure, of classes of formal languages, 192
 13521 cluster ranking, 362
 13523 clustering, 96
 K-means, 96
 exchange, 333
 hierarchical, 332
 soft, 97
 13528 co-hypernymy, 281
 13529 co-training, 108
 13530 code switching, 180, 186
 13531 Cohen's Kappa, 90
 13532 coherence, 396
 13533 cohesion, 379, 400
 13534 collective entity linking, 408
 13535 collocation extraction, 344
 13536 combinatorial categorial grammar, 220
 13537 complement clause, 213
 13538 complement event (probability), 476
 13539 composition (CCG), 221
 13540 compositional vector grammars, 391
 13541 compositionality, 2, 7, 10, 291, 341
 13542 computation graph, 48, 55
 dynamic, 57
 13544 computational linguistics (versus
 natural language processing), 1
 13546 computational social science, 5

- 13547 concept (in Abstract Meaning Representation), 319
 13549 conditional independence, 154, 480
 13550 conditional random field, 162
 13551 confidence interval, 86
 13552 configuration (transition-based parsing), 269
 13554 consistency, in logic, 291
 13555 constants, in logic, 286
 13556 constituents, 212
 13557 split, 312
 13558 constrained optimization, 315
 13559 content selection, 457
 13560 content words, 178
 13561 context-free grammars, 208
 13562 probabilistic (PCFGs), 235
 13563 synchronous, 441
 13564 weighted, 218, 227, 233
 13565 context-free languages, 207, 208
 13566 context-free production, 209
 13567 recursive, 209
 13568 unary, 210
 13569 context-sensitive languages, 218
 13570 continuous bag-of-words (CBOW), 334
 13571 contradiction, 346
 13572 conversational turns, 187
 13573 convexity, 29, 58, 300, 482, 485
 13574 biconvexity, 102
 13575 convolution
 13576 dilated, 64, 185
 13577 narrow, 63
 13578 one-dimensional, 63
 13579 wide, 63
 13580 convolutional neural networks, 62
 13581 in machine translation, 447
 13582 in relation extraction, 415
 13583 in semantic role labeling, 318
 13584 in sequence labeling, 170
 13585 cooperative principle, 351
 13586 coordinate ascent, 102
 13587 coordinating conjunctions, 178
 13588 copula, 177, 217, 260
 13589 coreference resolution, 347, 351
 13590 cross-document, 406
 13591 coreferent, 351
 13592 cosine similarity, 339, 380
 13593 cost function, 31
 13594 cost-augmented decoding, 32
 13595 coverage loss, 465
 13596 critical point, 58, 485
 13597 cross-entropy, 53, 416
 13598 cross-serial dependencies, 219
 13599 cross-validation, 24
 13600 crowdsourcing, 91
 13601 cumulative probability distribution, 86
 13602 decidability, 291
 13603 decision trees, 48
 13604 decoding
 13605 cost-augmented, 161
 13606 in conditional random fields, 163
 13607 definiteness, 183
 13608 delta function, 22
 13609 denotation, 286
 13610 dependency, 258
 13611 grammar, 257
 13612 graph, 258
 13613 labels, 259
 13614 path, 75, 314, 413
 13615 syntactic, 258
 13616 dependency parsing, 257
 13617 arc-eager, 269, 270
 13618 arc-factored, 263
 13619 arc-standard, 269
 13620 pseudo-projective, 272
 13621 second-order, 263
 13622 third-order, 264
 13623 derivation
 13624 in context-free languages, 209
 13625 in dependency parsing, 269
 13626 in semantic parsing, 293
 13627 determiner, 179

- | | | | |
|-------|-------------------------------------|-------|---|
| 13628 | phrase, 215 | 13668 | elementwise nonlinearity, 50 |
| 13629 | development set, 23, 81 | 13669 | Elman unit, 135 |
| 13630 | dialogue acts, 90, 187, 471 | 13670 | ELMo (embeddings from language models), 340 |
| 13631 | dialogue management, 468 | 13672 | embedding, 167 |
| 13632 | dialogue systems, 125, 466 | 13673 | emotion, 71 |
| 13633 | agenda-based, 468 | 13674 | empirical Bayes, 116 |
| 13634 | mixed-initiative, 468 | 13675 | empty string, 192 |
| 13635 | digital humanities, 5, 69 | 13676 | encoder-decoder model, 345, 442, 460 |
| 13636 | Dirichlet Compound Multinomial, 398 | 13677 | ensemble learning, 48, 318, 444 |
| 13637 | Dirichlet distribution, 115 | 13678 | entailment, 290, 346 |
| 13638 | discounting, 130 | 13679 | entity, 403 |
| 13639 | absolute, 130 | 13680 | embeddings, 407 |
| 13640 | discourse, 379 | 13681 | grid, 383 |
| 13641 | connectives, 385 | 13682 | linking, 351, 403, 405, 416 |
| 13642 | depth, 394 | 13683 | linking, collective, 409 |
| 13643 | parsing, 385 | 13684 | nil, 405 |
| 13644 | segment, 379 | 13685 | entropy, 42, 99 |
| 13645 | unit, 389 | 13686 | estimation, 482 |
| 13646 | discourse relations, 347, 384 | 13687 | EuroParl corpus, 435 |
| 13647 | coordinating, 389 | 13688 | evaluation |
| 13648 | implicit, 385 | 13689 | extrinsic, 139, 338 |
| 13649 | sense classification, 386 | 13690 | intrinsic, 139, 338 |
| 13650 | subordinating, 389 | 13691 | event, 403 |
| 13651 | distant supervision, 118, 418, 419 | 13692 | coreference, 421 |
| 13652 | distributed semantics, 327 | 13693 | detection, 421 |
| 13653 | distributional | 13694 | event semantics, 305 |
| 13654 | hypothesis, 325, 326 | 13695 | events, in probability, 475 |
| 13655 | semantics, 10, 327 | 13696 | disjoint, 476 |
| 13656 | statistics, 75, 248, 326 | 13697 | evidentiality, 182, 423 |
| 13657 | document frequency, 407 | 13698 | expectation, 481 |
| 13658 | domain adaptation, 95, 111 | 13699 | expectation-maximization, 97 |
| 13659 | by projection, 112 | 13700 | hard, 102 |
| 13660 | dropout, 57, 137 | 13701 | in language modeling, 131 |
| 13661 | dual decomposition, 317 | 13702 | in machine translation, 439 |
| 13662 | dynamic programming, 149 | 13703 | incremental, 102 |
| 13663 | E-step, 99 | 13704 | online, 102 |
| 13664 | early stopping, 27, 61 | 13705 | explicit semantic analysis, 328 |
| 13665 | early update, 276 | 13706 | factoid question, 424 |
| 13666 | edit distance, 201, 435 | 13707 | factor graph, 163 |
| 13667 | effective counts, 129 | | |

- 13708 factuality, 423
13709 fairness and bias, 5
13710 in machine translation, 434
13711 in word embeddings, 340
13712 false discovery rate, 87
13713 false negative, 81
13714 rate, 478
13715 false positive, 81, 479
13716 rate, 83, 478
13717 feature
13718 co-adaptation, 58
13719 function, 14
13720 hashing, 80
13721 noising, 58
13722 selection, 41
13723 features, 6
13724 blexical, 266
13725 collocation, 75
13726 emission, 148
13727 lexical, 47
13728 offset, 15
13729 pivot, 113
13730 transition, 148
13731 finite state
13732 acceptor, 193
13733 acceptor, chain, 205
13734 acceptors, weighted, 197
13735 automata, 193
13736 automaton, deterministic, 194
13737 composition, 204
13738 transducers, 196, 201
13739 finite-state
13740 transduction, 192
13741 fluency, 125, 433
13742 formal language theory, 191
13743 forward
13744 recurrence, 164
13745 variable, 164, 166
13746 forward-backward algorithm, 165, 206,
13747 240
13748 forward-looking centers, 382
13749 frame, 310
13750 element, 310
13751 in dialogue systems, 467
13752 FrameNet, 310
13753 Frobenius norm, 57, 105
13754 function words, 178
13755 function, in logic, 289
13756 functional segmentation, 379, 381
13757 garden path sentence, 146
13758 gazetteer, 184, 357, 413
13759 generalization, 27
13760 of neural networks, 59
13761 generalized linear models, 42
13762 generative model
13763 for classification, 17
13764 for coreference, 363
13765 for interpolated language modeling,
13766 131
13767 for parsing, 235
13768 for sequence labeling, 154
13769 Gibbs sampling, 115, 409
13770 collapsed, 116
13771 gloss, 125, 179, 433
13772 government and binding theory, 353
13773 gradient, 29
13774 clipping, 60
13775 descent, 38
13776 exploding, 137
13777 vanishing, 51, 137
13778 Gram matrix, 414
13779 grammar equivalence, 210
13780 grammar induction, 241
13781 grammaticality, 396
13782 graphical model, 154
13783 graphics processing units (GPUs), 170,
13784 185
13785 grid search, 23
13786 Hamming cost, 161
13787 Hansards corpus, 435

- 13788 hanzi, 77
 13789 head rules, for lexicalized CFGs, 246, 257
 13790 head word, 212, 244, 257, 353
 of a dependency edge, 258
 13792 hedging, 423
 13793 held-out data, 139
 13794 Hessian matrix, 39
 13795 hidden Markov models, 154
 13796 hierarchical recurrent network, 470
 13797 highway network, 52
 13798 holonymy, 75
 13799 homonymy, 73
 13800 human computation, 91
 13801 hypergraph, 393
 13802 hypernymy, 75, 281
 13803 hyperparameter, 22
 13804 hyponymy, 75
 13805 illocutionary force, 187
 13806 importance sampling, 453
 13807 importance score, 453
 13808 independent and identically distributed
 (IID), 17
 13809 inference
 in structured prediction, 147
 13812 logical, 285
 13813 rules for propositional logic, 288
 13814 inflection point, 486
 13815 information extraction, 403
 13816 open, 419
 13817 information retrieval, 5, 416
 13818 inside recurrence, 235–237
 13819 inside-outside algorithm, 240, 249
 13820 instance labels, 16
 13821 instance, in Abstract Meaning
 Representation, 319
 13823 integer linear programming, 315
 13824 in coreference resolution, 362
 13825 in entity linking, 409
 13826 in extractive summarization, 395
 13827 in sentence compression, 466
 13828 inter-annotator agreement, 90
 13829 interjections, 177
 13830 interlingua, 432
 13831 interpolation, 131, 199
 13832 interval algebra, 421
 13833 inverse document frequency, 407
 13834 inversion (of finite state automata), 203
 13835 irrealis, 70
 13836 Jensen’s inequality, 99
 13837 Kalman smoother, 172
 13838 kernel, 48, 414
 13839 Kleene star, 192
 13840 knapsack problem, 395
 13841 knowledge base, 403
 13842 population, 416
 13843 label bias problem, 275, 280
 13844 label propagation, 110, 120
 13845 Lagrangian, 487
 13846 lambda calculus, 292
 13847 lambda expressions, 292
 13848 language model, 4, 126
 13849 latent Dirichlet allocation, 380
 13850 latent semantic analysis, 328, 329
 13851 latent variable, 98, 206, 299, 419, 432
 13852 conditional random field, 301
 13853 in parsing, 249
 13854 perceptron, 207, 300, 360
 13855 layer normalization, 61, 447
 13856 learning
 active, 118
 13858 batch, 25
 13859 constraint-driven, 118
 13860 deep, 47
 13861 discriminative, 25
 13862 multiple instance, 118, 418
 13863 multitask, 118
 13864 online, 25, 39
 13865 reinforcement, 365, 451
 13866 semi-supervised, 76, 95, 339

- | | | | |
|-------|--|-------|-------------------------------------|
| 13867 | to search, 257, 277, 366 | 13908 | memory cell, 137 |
| 13868 | transfer, 118 | 13909 | lookup layer, 53, 135 |
| 13869 | unsupervised, 71, 95 | 13910 | loss |
| 13870 | learning rate, 38 | 13911 | function, 27 |
| 13871 | least squares, 72 | 13912 | hinge, 29 |
| 13872 | leave-one-out cross-validation, 24 | 13913 | logistic, 36 |
| 13873 | lemma, 73, 202 | 13914 | WARP, 411 |
| 13874 | lemmatization, 79 | 13915 | zero-one, 28 |
| 13875 | lexical entry, 293 | 13916 | machine learning, 2 |
| 13876 | lexical unit, 310 | 13917 | supervised, 16 |
| 13877 | lexicalization | 13918 | theory, 26 |
| 13878 | in parsing, 245 | 13919 | machine reading, 425 |
| 13879 | in text generation, 457 | 13920 | macro, 404 |
| 13880 | lexicalized tree-adjoining grammar for | 13921 | micro, 404 |
| 13881 | discourse (D-LTAG), 385 | 13922 | machine translation, 125 |
| 13882 | lexicon, 293 | 13923 | neural, 432 |
| 13883 | in combinatory categorial grammar, | 13924 | statistical, 432 |
| 13884 | 221 | 13925 | margin, 26, 30 |
| 13885 | lexicon-based classification, 72 | 13926 | functional, 32 |
| 13886 | seed, 73 | 13927 | geometric, 32 |
| 13887 | light verb, 321 | 13928 | marginal relevance, 465 |
| 13888 | linear separability, 25 | 13929 | marginalization, 477 |
| 13889 | linearization, 466, 472 | 13930 | markable, 358 |
| 13890 | link function, 42 | 13931 | Markov assumption, 154 |
| 13891 | literal character, 192 | 13932 | Markov blanket, 154 |
| 13892 | local optimum, 102, 486 | 13933 | Markov Chain Monte Carlo (MCMC), |
| 13893 | locally-normalized objective, 275 | 13934 | 103, 115, 172 |
| 13894 | log-bilinear language model, 342 | 13935 | Markov decision process, 468 |
| 13895 | log-likelihood | 13936 | partially-observable (POMDP), 470 |
| 13896 | conditional, 52 | 13937 | Markov random fields, 162 |
| 13897 | log-linear models, 42 | 13938 | Markovization |
| 13898 | logic, 287 | 13939 | vertical, 244 |
| 13899 | first-order, 288 | 13940 | matrix-tree theorem, 268 |
| 13900 | higher-order, 289 | 13941 | max-margin Markov network, 161 |
| 13901 | propositional, 287 | 13942 | max-product algorithm, 157 |
| 13902 | logistic function, 42 | 13943 | maximum a posteriori, 22, 483 |
| 13903 | long short-term memory (LSTM), 52, | 13944 | maximum conditional likelihood, 36 |
| 13904 | 135, 137, 181, 442 | 13945 | maximum directed spanning tree, 264 |
| 13905 | bidirectional, 169, 445 | 13946 | maximum entropy, 42 |
| 13906 | deep, 443 | 13947 | maximum likelihood, 17, 21, 482 |
| 13907 | LSTM-CRF, 169, 318 | | |

- 13948 McNemar’s test, 84
 13949 meaning representation, 285
 13950 membership problem, 191
 13951 mention
 in coreference resolution, 351
 in entity linking, 403
 in information extraction, 405
 13955 mention ranking, 360
 13956 mention-pair model, 359
 13957 meronymy, 75
 13958 method of moments, 117
 13959 mildly context-sensitive languages, 219
 13960 minibatch, 39
 13961 minimization of finite-state automata,
 196
 13963 minimum error-rate training (MERT),
 452
 13965 minimum risk training, 452
 13966 modality, 422
 13967 model, 8, 38
 13968 modifier, 258
 13969 modus ponens, 288, 303
 13970 moment-matching, 42
 13971 monomorphemic, 196
 13972 morpheme, 6, 141, 195
 13973 morphological segmentation, 159
 13974 morphology, 79, 158, 194, 342, 448
 derivational, 194
 inflectional, 177, 194, 202
 13977 morphosyntactic, 176
 attributes, 180
 13979 morphotactics, 195
 13980 multi-view learning, 108
 13981 multiclass classification
 one-versus-all, 414
 one-versus-one, 414
 13984 multinomial
 distribution, 18
 13986 Naïve Bayes, 19
 13987 Naïve Bayes, 17
 13988 name dictionary, 406
 13989 named entity, 183
 linking, 405
 13991 recognition, 169, 403, 405
 13992 types, 405
 13993 nearest-neighbor, 48, 414
 13994 negation, 70, 422
 scope, 423
 13996 negative sampling, 336, 337, 411
 13997 Neo-Davidsonian event semantics, 306
 13998 neural attention, 371, 426, 443, 444, 460
 coarse-to-fine, 463
 structured, 463
 14001 neural gate, 52, 445
 14002 neural network, 48
 adversarial, 113
 14004 bidirectional recurrent, 168
 14005 convolutional, 53, 62, 185, 415
 14006 feedforward, 50
 14007 recurrent, 134, 415
 14008 recursive, 250, 343, 346, 396
 14009 noise-contrastive estimation, 135
 14010 noisy channel model, 126, 436
 14011 nominal modifier, 215
 14012 nominals, 351, 357
 14013 non-terminals, 209
 14014 normalization, 78
 14015 noun phrase, 2, 212
 14016 nouns, 176
 14017 NP-hard, 41, 409
 14018 nucleus, in RST, 389
 14019 null hypothesis, 84
 14020 null subjects, 375
 14021 numerals, 179
 14022 one-hot vector, 53
 14023 online support vector machine, 31
 14024 ontology, 9
 14025 open word classes, 176
 14026 opinion polarity, 69
 14027 optimization

- | | | | |
|-------|--|-------|--|
| 14028 | batch, 38 | 14068 | phonology, 196 |
| 14029 | combinatorial, 8 | 14069 | phrase, 212 |
| 14030 | constrained, 33 | 14070 | phrase-structure grammar, 212 |
| 14031 | convex, 38 | 14071 | pointwise mutual information, 281, 329,
330 |
| 14032 | numerical, 8 | 14072 | in collocation identification, 344 |
| 14033 | quasi-Newton, 39 | 14073 | positive, 331 |
| 14034 | oracle, 275 | 14075 | shifted positive, 337 |
| 14035 | dynamic, 277 | 14076 | policy, 276, 365, 469 |
| 14036 | in learning to search, 366 | 14077 | policy gradient, 366 |
| 14037 | orthography, 196, 204 | 14078 | Polysemy, 74 |
| 14038 | orthonormal matrix, 60 | 14079 | pooling, 63, 65, 372, 460 |
| 14039 | out-of-vocabulary words, 181 | 14080 | positional encodings |
| 14040 | outside recurrence, 237, 240 | 14081 | in machine translation, 447 |
| 14041 | overfitting, 23, 27 | 14082 | in relation extraction, 415 |
| 14042 | in neural networks, 59 | 14083 | power law, 2 |
| 14043 | overgeneration, 203, 213 | 14084 | pragmatics, 351 |
| 14044 | parallel corpus, 435 | 14085 | precision, 82, 478 |
| 14045 | parameters, 482 | 14086 | at- k , 83, 398 |
| 14046 | paraphrase, 346 | 14087 | labeled, 230 |
| 14047 | parent annotation, 244 | 14088 | unlabeled, 230 |
| 14048 | parsing, 209 | 14089 | precision-recall curve, 83, 418 |
| 14049 | agenda-based, 279 | 14090 | predicate, 403 |
| 14050 | chart, 226 | 14091 | predicative adjectives, 217 |
| 14051 | easy-first, 279 | 14092 | predictive likelihood, 103 |
| 14052 | graph-based, 263 | 14093 | prepositional phrase, 2, 217 |
| 14053 | transition-based, 225 | 14094 | primal form, 487 |
| 14054 | part-of-speech, 6, 175 | 14095 | prior expectation, 483 |
| 14055 | tagging, 145 | 14096 | probabilistic models, 482 |
| 14056 | particle, 179, 217 | 14097 | probabilistic topic model, 5, 409 |
| 14057 | partition, 477 | 14098 | probability |
| 14058 | partition function, 164 | 14099 | chain rule, 477 |
| 14059 | passive-aggressive, 43, 487 | 14100 | conditional, 35, 477, 481 |
| 14060 | path, in an FSA, 193 | 14101 | density function, 481 |
| 14061 | Penn Discourse Treebank (PDTB), 385 | 14102 | distribution, 480 |
| 14062 | Penn Treebank, 140, 159, 180, 212, 238 | 14103 | joint, 17, 35, 481 |
| 14063 | perceptron, 25 | 14104 | likelihood, 478 |
| 14064 | incremental, 277, 364 | 14105 | marginal, 481 |
| 14065 | multilayer, 50 | 14106 | mass function, 85, 481 |
| 14066 | structured, 160 | 14107 | posterior, 478 |
| 14067 | perplexity, 140 | 14108 | prior, 22, 36, 478 |

- 14109 simplex, 18
 14110 processes, 422
 14111 productivity, 195
 14112 projectivity, 261
 14113 pronominal anaphora resolution, 351
 14114 pronoun, 178
 reflexive, 353
 14115 PropBank, 310
 14116 proper nouns, 177
 14118 proposal distribution, 453
 14119 propositions, 286, 287, 422
 14120 prosody, 187
 14121 proto-roles, 309
 14122 pumping lemma, 207
 14123 pushdown automata, 209, 251
 14124 quadratic program, 34
 14125 quantifier, 289
 existential, 289
 universal, 289
 14128 question answering, 346, 405
 cloze, 425
 extractive, 426
 multiple-choice, 425
 14132 random outcomes, 475
 14133 random variable, 480
 discrete, 480
 indicator, 480
 14136 ranking, 406
 loss, 406
 14138 recall, 82, 478
 14139 at- k , 398
 14140 labeled, 230
 14141 unlabeled, 230
 14142 receiver operating characteristic (ROC), 83
 14144 rectified linear unit (ReLU), 51
 leaky, 51
 14146 reference arguments, 322
 14147 reference resolution, 351
 14148 reference translations, 433
 14149 referent, 351
 generic, 356
 14151 referring expressions, 351, 381, 458
 14152 regression, 72
 linear, 72
 logistic, 35
 ridge, 72
 14156 regular expression, 192
 14157 regular language, 192
 14158 regularization, 23, 32
 14159 reification (events), 305
 14160 relation
 extraction, 277
 logical, 286, 295
 14163 relation extraction, 411
 14164 relations
 in information extraction, 403
 14166 relative frequency estimate, 21, 126, 483
 14167 reranking, 250
 14168 residual networks, 52
 14169 retrofitting, 343
 14170 Rhetorical Structure Theory (RST), 389
 14171 rhetorical zones, 381
 14172 risk, 452
 14173 roll-in, roll-out, 366
 14174 saddle point, 58, 486
 14175 sample space, 475
 14176 satellite, in RST, 389
 14177 satisfaction, 290
 14178 scheduled sampling, 451
 14179 schema, 403, 404, 419
 14180 search error, 252, 363
 14181 segmented discourse representation
 theory (SDRT), 384
 14183 self-attention, 446
 14184 self-training, 108
 14185 semantic concordance, 76
 14186 semantic parsing, 291
 14187 semantic role, 306

- 14188 semantic role labeling, 306, 412, 420
 14189 semantics, 285
 dynamic, 302, 384
 14190 in parsing, 242
 14192 lexical, 73
 14193 model-theoretic, 286
 14194 underspecification, 302
 14195 semi-supervised learning, 105
 14196 semiring, 199
 algebra, 172
 14198 expectation, 207
 14199 tropical, 173
 14200 sentence compression, 465
 14201 sentence fusion, 465
 14202 sentence, in logic, 289
 14203 sentiment analysis, 69
 lexicon-based, 70
 targeted, 71
 14206 sentiment lexicon, 16
 14207 sequence-to-sequence model, 442
 14208 shift-reduce parsing, 251
 14209 shortest-path algorithm, 197
 14210 sigmoid, 49
 14211 simplex, 115
 14212 singular value decomposition, 60, 104
 truncated, 104, 330
 14214 singular vectors, 60
 14215 skipgram word embeddings, 335
 14216 slack variables, 34
 14217 slot filling, 416
 14218 slots, in dialogue systems, 467
 14219 smooth functions, 29
 14220 smoothing, 22, 129
 Jeffreys-Perks, 129
 Kneser-Ney, 133
 Laplace, 22, 129
 Lidstone, 129
 14225 softmax, 49, 134, 416
 hierarchical, 135, 336
 14227 source language, 431
 14228 spanning tree, 258
 sparse matrix, 330
 sparsity, 41
 14231 spectral learning, 117
 14232 speech acts, 187
 14233 speech recognition, 125
 14234 squashing function, 51, 135
 14235 stand-off annotations, 90
 14236 Stanford Natural Language Inference corpus, 346
 14237 statistical significance, 84
 14239 stemming, 7, 78, 195
 14240 step size, 486
 14241 stochastic gradient descent, 29, 35, 39
 14242 stopwords, 80
 14243 string, in formal language theory, 191
 14244 string-to-tree translation, 441
 14245 strong compositionality criterion, 390
 14246 structure induction, 170
 14247 structure prediction, 15
 14248 subgradient, 29, 41
 14249 subjectivity detection, 70
 14250 subordinating conjunctions, 178
 14251 sum-product algorithm, 164
 14252 summarization, 125, 393
 abstractive, 393, 464
 extractive, 393
 multi-document, 465
 of sentences, 464
 14255 supersenses, 339
 14258 support vector machine, 34
 kernel, 48, 414
 structured, 161
 14261 support vectors, 34
 14262 surface form, 203
 14263 surface realization, 457
 14264 synonymy, 74, 281, 325
 14265 synset, 74, 369
 14266 syntactic path, 313
 14267 syntactic-semantic grammar, 293
 14268 syntax, 175, 211, 285

- 14269 tagset, 176
 14270 tanh activation function, 51
 14271 target language, 431
 14272 tense, 177
 14273 terminal symbols, 209
 14274 test set, 23, 95
 14275 test statistic, 84
 14276 text mining, 5
 14277 text planning, 457
 14278 thematic roles, 307
 14279 third axiom of probability, 476
 14280 TimeML, 421
 14281 tokenization, 77, 185
 14282 tokens and types, 19
 14283 topic segmentation, 379
 hierarchical, 381
 14284 trace, 222
 14285 training set, 17, 95
 14287 transformer architecture, 446
 14288 transition system, 269
 for context-free grammars, 251
 14289 transitive closure, 361
 14291 translation error rate (TER), 435
 14292 translation model, 126
 14293 transliteration, 449
 14294 tree-adjoining grammar, 220
 14295 tree-to-string translation, 442
 14296 tree-to-tree translation, 441
 14297 treebank, 238
 14298 trellis, 150, 205
 14299 trigrams, 24
 14300 trilexical dependencies, 248
 14301 tropical semiring, 199
 14302 true negative, 81
 14303 true positive, 81, 479
 rate, 83
 14305 truth conditions, 290
 14306 tuning set, *see* development set, 81
 14307 Turing test, 3
 14308 two-tailed test, 86
 14309 type systems, 295
- 14310 type-raising, 221, 295
 14311 unary closure, 228
 14312 undercutting, in argumentation, 393, 399
 14313 underfitting, 23
 14314 underflow, 17
 14315 undergeneration, 203, 213
 14316 Universal Dependencies, 176, 257
 14317 unseen word, 169
 14318 utterances, 187
- 14319 validation function, 301
 14320 validity, in logic, 290
 14321 value iteration, 469
 14322 variable, 319
 bound, 289
 free, 289
 14325 variance, 22, 87
 14326 Vauquois Pyramid, 432
 14327 verb phrase, 213
 14328 VerbNet, 308
 14329 verbs, 177
 14330 Viterbi
 algorithm, 148
 variable, 149
 14333 volition, 307
- 14334 weight decay, 57
 14335 Wikification, 405
 14336 Winograd schemas, 3
 14337 word
 embeddings, 47, 53, 135, 136, 327,
 334
 embeddings, fine-tuned, 340
 embeddings, pre-trained, 339
 14342 representations, 326
 14343 sense disambiguation, 73
 14344 senses, 73, 309
 14345 tokens, 77
 14346 world model, 286
 14347 builder, 291
 14348 checker, 291

14349 yield, 209

14350 Zipf's law, 143