# Natural Language Processing

Jacob Eisenstein

January 28, 2018

# Contents

# Preface

This text is built from the notes that I use for teaching Georgia Tech's undergraduate and graduate courses on natural language processing, CS 4650 and 7650. There are several other good resources (e.g., Manning and Schütze, 1999; Jurafsky and Martin, 2009; Smith, 2011; Figueiredo et al., 2013; Collins, 2013), but for various reasons I wanted to create something of my own.

The text assumes familiarity with basic linear algebra, and with calculus through Lagrange multipliers. It includes a refresher on probability, but some previous exposure would be helpful. An introductory course on the analysis of algorithms is also assumed; in particular, the reader should be familiar with asymptotic analysis of the time and memory costs of algorithms, and should have seen dynamic programming. No prior background in machine learning or linguistics is assumed, and even students with background in machine learning should be sure to read the introductory chapters, since the notation used in natural language processing is different from typical machine learning presentations, due to the emphasis on structure prediction in applications of machine learning to language. Throughout the book, advanced material is marked with an asterisk, and can be safely skipped.

The notes focus on what I view as a core subset of the field of natural language processing, unified by the concepts of linear models and structure prediction. A remarkable thing about the field of natural language processing is that so many problems in language technology can be solved by a small number of methods. These notes focus on the following methods:

**Search algorithms** shortest path, Viterbi, CKY, minimum spanning tree, shift-reduce, integer linear programming, beam search.

**Learning algorithms** Naïve Bayes, logistic regression, perceptron, expectation-maximization, matrix factorization, backpropagation.

The goal of this text is to teach how these methods work, and how they can be applied to problems that arise in the computer processing of natural language: document classification, word sense disambiguation, sequence labeling (part-of-speech tagging and named

entity recognition), parsing, coreference resolution, relation extraction, discourse analysis, and, to a limited degree, language modeling and machine translation. Proper application of these techniques requires understanding the underlying linguistic phenomena, and the notes therefore include a minimal foundation in morphology, syntax, semantics, and discourse. However, a detailed understanding of these topics can be provided only by a linguistics textbook (e.g., Akmajian et al., 2010; Fromkin et al., 2013).

-Jacob Eisenstein, January 28, 2018

# Notation

As a general rule, random variables and observable counts are indicated with Roman letters ($a, b, c$), and parameters are indicated with Greek letters ($\alpha, \beta, \theta$). Vectors are indicated with bold script for both random variables $\boldsymbol{x}$ 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, $\log_2 x$ |

| **Linear algebra** | |
|---|---|
| $\boldsymbol{x}^{(i)}$ | a column vector of feature counts for instance $i$, often word counts |
| $\boldsymbol{x}_{j:k}$ | elements $j$ through $k$ (inclusive) of a vector $\boldsymbol{x}$ |
| $[\boldsymbol{x}; \boldsymbol{y}]$ | vertical concatenation of two column vectors |
| $[\boldsymbol{x}, \boldsymbol{y}]$ | horizontal concatenation of two column vectors |
| $\boldsymbol{e}_n$ | 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 $\boldsymbol{\theta}$ |
| $\boldsymbol{\theta} \cdot \boldsymbol{x}^{(i)}$ | the dot product of vectors $\boldsymbol{\theta}$ and $\boldsymbol{x}^{(i)}$ |
| $\mathbf{X}$ | a matrix |
| $x_{i,j}$ | row $i$, column $j$ 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 $K = \lvert\mathcal{Y}\rvert$ |

| **Probabilities** | |
|---|---|
| $\Pr(A)$ | probability of event $A$ |
| $\Pr(A \mid B)$ | probability of event $A$, conditioned on event $B$ |
| $\mathrm{p}_B(b)$ | the marginal probability of random variable $B$ taking value $b$; written $\mathrm{p}(b)$ when the choice of random variable is clear from context |
| $\mathrm{p}_{B\mid A}(b \mid a)$ | the probability of random variable $B$ taking value $b$, conditioned on $A$ taking value $a$; written $\mathrm{p}(b \mid a)$ when clear from context |
| $a \mid b \sim f(\theta_b)$ | conditioned on $b$, $a$ is drawn from distribution of form $f$ with parameters $\theta_b$ [todo: decide about upper-case random variables] |

| **Linear classification** | |
|---|---|
| $\boldsymbol{f}(\boldsymbol{x}^{(i)}, y)$ | feature vector for instance $i$ with label $y$ |
| $\boldsymbol{\theta}$ | 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 |
| $\lambda$ | the amount of regularization |

| **Sequence labeling** | |
|---|---|
| $\mathcal{Y}(\boldsymbol{w})$ | the set of possible tag sequences for the word sequence $\boldsymbol{w}$ |
| $\lozenge$ | the start tag |
| $\blacklozenge$ | the stop tag |
| $\square$ | the start token |
| $\blacksquare$ | the stop token |

# Part I

# Words, bags of words, and features

# Chapter 1

# Linear text classification

We'll start with the problem of **text classification**: given a text document, assign it a discrete label $y \in \mathcal{Y}$, where $\mathcal{Y}$ is the set of possible labels. This problem has many applications, from spam filtering to analysis of electronic health records. Text classification is also a building block that is used throughout more complex natural language processing tasks.

To perform this task, the first question is how to represent each document. A common approach is to use a vector of word counts, e.g., $\boldsymbol{x}^{(i)} = [0\ 1\ 1\ 0\ 0\ 2\ 0\ 1\ 13\ 0\ \ldots]^\top$, where $x_j^{(i)}$ is the count of word $j$ in document $i$. The length of $\boldsymbol{x}^{(i)}$ is $V \triangleq |\mathcal{V}|$, where $\mathcal{V}$ is the set of possible words in the vocabulary.

The object $\boldsymbol{x}^{(i)}$ is a vector, but colloquially we call it a **bag of words**, because it includes only information about the count of each word, and not the order in which the words appear. We have thrown out grammar, sentence boundaries, paragraphs — everything but the words. Yet the bag of words model is surprisingly effective for text classification. If you see the word *freeee* in an email, is it a spam email? What if you see the word *Bayesian*? For many labeling problems, individual words can be strong predictors.

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. In the spam filtering case, we might assign a positive score to the word *freeee* for the label SPAM, and a negative score to the word *Bayesian*. These scores are called **weights**, and they are arranged in a column vector $\boldsymbol{\theta}$.

Suppose that you want a multiclass classifier, where $K \triangleq |\mathcal{Y}| > 2$. For example, we might want to classify news stories about sports, celebrities, music, and business. The goal is to predict a label $\hat{y}^{(i)}$, given the bag of words $\boldsymbol{x}^{(i)}$, using the weights $\boldsymbol{\theta}$. For each label $y \in \mathcal{Y}$, we compute a score $\psi(\boldsymbol{x}^{(i)}, y)$, which is a scalar measure of the compatibility between the bag-of-words $\boldsymbol{x}^{(i)}$ and the label $y$. In a linear bag-of-words classifier, this score

is the vector inner product between the weights $\boldsymbol{\theta}$ and the output of a **feature function** $\boldsymbol{f}(\boldsymbol{x}^{(i)}, y)$,

$$\psi(\boldsymbol{x}^{(i)}, y) = \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y). \quad [1.1]$$

As the notation suggests, the $\boldsymbol{f}$ is a *function* of two arguments, the word counts $\boldsymbol{x}^{(i)}$ and the label $y$, and it returns a vector output. For example, for arguments $\boldsymbol{x}^{(i)}$ and $y$, element $j$ of this feature vector might be,

$$f_j(\boldsymbol{x}, y) = \begin{cases} x_{\textit{freeee}}, & \text{if } y = \textsc{Spam} \\ 0, & \text{otherwise} \end{cases} \quad [1.2]$$

This function returns the count of the word *freeee* if the label is SPAM, and it returns zero otherwise. The corresponding weight $\theta_j$ then scores the compatibility of the word *freeee* with the label SPAM. A positive score means that this word makes the label more likely.

To formalize this feature function, we define $\boldsymbol{f}(\boldsymbol{x}, y)$ as,

$$\boldsymbol{f}(\boldsymbol{x}, y = 0) = [\boldsymbol{x}; \underbrace{0; 0; \ldots; 0}_{(K-1)\times V}] \quad [1.3]$$

$$\boldsymbol{f}(\boldsymbol{x}, y = 1) = [\underbrace{0; 0; \ldots; 0}_{V}; \boldsymbol{x}; \underbrace{0; 0; \ldots; 0}_{(K-2)\times V}] \quad [1.4]$$

$$\boldsymbol{f}(\boldsymbol{x}, y = 2) = [\underbrace{0; 0; \ldots; 0}_{2\times V}; \boldsymbol{x}; \underbrace{0; 0; \ldots; 0}_{(K-3)\times V}] \quad [1.5]$$

$$\boldsymbol{f}(\boldsymbol{x}, y = K) = [\underbrace{0; 0; \ldots; 0}_{(K-1)\times V}; \boldsymbol{x}], \quad [1.6]$$

where $[\underbrace{0; 0; \ldots; 0}_{(K-1)\times V}]$ is a column vector of $(K-1) \times V$ zeros, and the semicolon indicates vertical concatenation. This arrangement is shown in Figure 1.1; the notation may seem awkward at first, but it generalizes to an impressive range of learning settings.

Given a vector of weights, $\boldsymbol{\theta} \in \mathbb{R}^{V \times K}$, we can now compute the score $\psi(\boldsymbol{x}, y)$. This inner product gives a scalar measure of the compatibility of the observation $\boldsymbol{x}$ with label $y$.[1] For any document $\boldsymbol{x}^{(i)}$, we predict the label $\hat{y}$,

$$\hat{y} = \underset{y \in \mathcal{Y}}{\text{argmax}}\, \psi(\boldsymbol{x}^{(i)}, y) \quad [1.7]$$

$$\psi(\boldsymbol{x}^{(i)}, y) = \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y). \quad [1.8]$$

[1] Only $V \times (K-1)$ features and weights are necessary: by stipulating that $\psi(\boldsymbol{x}, y = K) = 0$ regardless of $\boldsymbol{x}$, 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 $y = \text{Sign}(\boldsymbol{\beta} \cdot \boldsymbol{x} + a)$, 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 $\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}, y)$ for all $y \in \mathcal{Y}$.

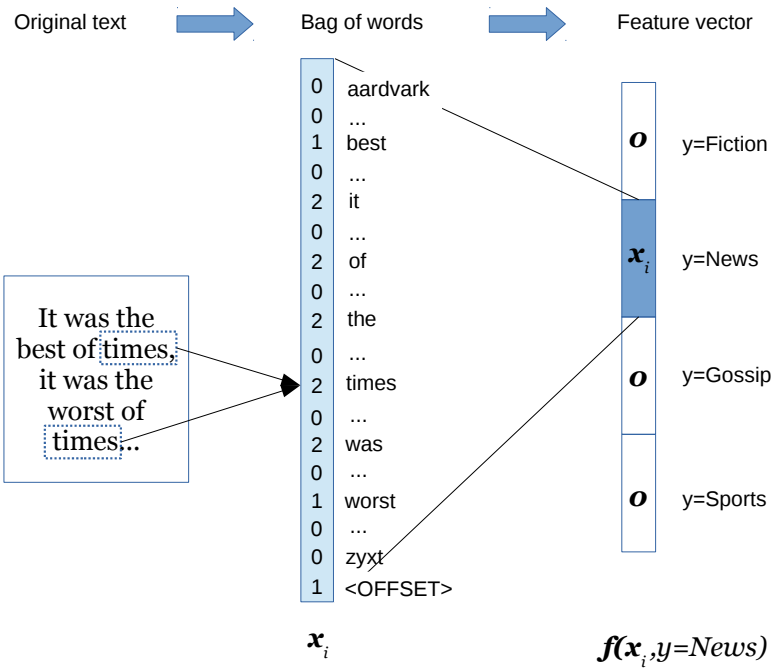



Figure 1.1: The bag-of-words and feature vector representations, for a hypothetical text classification task.

This inner product notation gives a clean separation between the *data* ($\boldsymbol{x}$ and $y$) and the *parameters* ($\boldsymbol{\theta}$). This notation also generalizes nicely to **structured prediction**, in which the space of labels $\mathcal{Y}$ is very large, and we want to model shared substructures between labels.

It is common to add an **offset feature** at the end of the vector of word counts $\boldsymbol{x}$, which is always 1. We then have to also add an extra zero to each of the zero vectors, to make the vector lengths match. This gives the entire feature vector $\boldsymbol{f}(\boldsymbol{x}, y)$ a length of $(V + 1) \times K$. The weight associated with this offset feature can be thought of as a bias for (or against) each label. For example, if we expect most documents to be spam, then the weight for the offset feature for $y = \text{SPAM}$ should be larger than the weight for the offset feature for $y = \text{HAM}$.

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

associated dictionary. For example,[2]

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

Similarly, if we want to distinguish positive and negative sentiment, we could use positive and negative **sentiment lexicons**, which are defined by social psychologists (Tausczik and Pennebaker, 2010).

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

## 1.1 Naïve Bayes

The **joint probability** of a bag of words $\boldsymbol{x}^{(i)}$ and its true label $y^{(i)}$ is written $\text{p}(\boldsymbol{x}^{(i)}, y^{(i)})$. Suppose we have a dataset of $N$ labeled instances, $\{\boldsymbol{x}^{(1:N)}, y^{(1:N)}\}$, which we assume are **independent and identically distributed (IID)** (see § A.3). Then the joint probability of the entire dataset is,

$$\text{p}(\boldsymbol{x}^{(1:N)}, y^{(1:N)}) = \prod_{i=1}^{N} \text{p}(\boldsymbol{x}^{(i)}, y^{(i)}). \tag{1.9}$$

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

$$\begin{aligned}
\boldsymbol{\theta} &= \operatorname*{argmax}_{\boldsymbol{\theta}} \text{p}(\boldsymbol{x}^{(1:N)}, y^{(1:N)}; \boldsymbol{\theta}) && [1.10] \\
&= \operatorname*{argmax}_{\boldsymbol{\theta}} \prod_{i=1}^{N} \text{p}(\boldsymbol{x}^{(i)}, y^{(i)}; \boldsymbol{\theta}) && [1.11] \\
&= \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^{N} \log \text{p}(\boldsymbol{x}^{(i)}, y^{(i)}; \boldsymbol{\theta}). && [1.12]
\end{aligned}$$

[2] In this notation, each tuple (language, word) indexes an element in $\boldsymbol{\theta}$, which remains a vector.

**Algorithm 1** Generative process for the Naïve Bayes classifier

**for** Document $i \in \{1, 2, \ldots, N\}$ **do**:
  Draw the label $y^{(i)} \sim \text{Categorical}(\boldsymbol{\mu})$;
  Draw the word counts $\boldsymbol{x}^{(i)} \mid y^{(i)} \sim \text{Multinomial}(\boldsymbol{\phi}_{y^{(i)}})$.

The notation $\text{p}(\boldsymbol{x}^{(i)}, y^{(i)}; \boldsymbol{\theta})$ indicates that $\boldsymbol{\theta}$ is a *parameter* of the probability function. We can move from a product of probabilities to a sum of log probabilities because the $\log$ function is monotonically increasing over positive arguments, and so the same $\boldsymbol{\theta}$ will maximize both the probability and the log probability. Working with logarithms is desirable because of numerical stability: on a large dataset, multiplying many probabilities can **underflow** to zero.[3]

The probability $\text{p}(\boldsymbol{x}^{(i)}, y^{(i)}; \boldsymbol{\theta})$ is defined through a **generative model** — an idealized random process that has generated the observed data, $\langle \boldsymbol{x}^{(i)}, y^{(i)} \rangle$.[4] Algorithm 1 describes the generative model describes the **Naïve Bayes** classifier, with parameters $\boldsymbol{\theta} = \{\boldsymbol{\mu}, \boldsymbol{\phi}\}$.

- The first line of this generative model encodes the assumption that the instances are mutually independent: neither the label nor the text of document $i$ affects the label or text of document $j$.[5] Furthermore, the instances are identically distributed: the distributions over the label $y^{(i)}$ and the text $\boldsymbol{x}^{(i)}$ (conditioned on $y^{(i)}$) are the same for all instances $i$.
- The second line of the generative model states that the random variable $y^{(i)}$ is drawn from a categorical distribution with parameter $\boldsymbol{\mu}$. Categorical distributions are like weighted dice: the vector $\boldsymbol{\mu} = [\mu_1, \mu_2, \ldots, \mu_K]$ gives the probabilities of each label, so that the probability of drawing label $y$ is equal to $\mu_y$. For example, if $\mathcal{Y} = \{\text{POSITIVE}, \text{NEGATIVE}, \text{NEUTRAL}\}$, we might have $\boldsymbol{\mu} = [0.1, 0.7, 0.2]$. We require $\sum_y \mu_y = 1$ and $\mu_y \geq 0, \forall_y$.[6]
- The third line describes how the bag-of-words counts $\boldsymbol{x}^{(i)}$ are generated. By writing $\boldsymbol{x}^{(i)} \mid y^{(i)}$, this line indicates that the word counts are conditioned on the label, so

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

[4] Generative 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).

[5] Can you think of any cases in which this assumption is too strong?

[6] Formally, we require $\boldsymbol{\mu} \in \Delta^{K-1}$, where $\Delta^{K-1}$ 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 $K$.

that the joint probability is factored using the chain rule,

$$\text{p}(\boldsymbol{x}^{(i)}, y^{(i)}) = \text{p}(\boldsymbol{x}^{(i)} \mid y^{(i)}) \times \text{p}(y^{(i)}). \quad [1.13]$$

The specific distribution $\text{p}_{\boldsymbol{x}^{(i)}|y^{(i)}}$ is the **multinomial**, which is a probability distribution over vectors of non-negative counts. The probability mass function for this distribution is:

$$\text{p}_{\text{mult}}(\boldsymbol{x}; \boldsymbol{\phi}) = B(\boldsymbol{x}) \prod_{j=1}^{V} \phi_j^{x_j} \quad [1.14]$$

$$B(\boldsymbol{x}) = \frac{\left(\sum_{j=1}^{V} x_j\right)!}{\prod_{j=1}^{V}(x_j!)} \quad [1.15]$$

As in the categorical distribution, the parameter $\phi_j$ can be interpreted as a probability: specifically, the probability that any given token in the document is the word $j$.[7] The multinomial distribution involves a product over words, with each term in the product equal to the probability $\phi_j$, exponentiated by the count $x_j$. Words that have zero count play no role in this product, because $\phi_j^0 = 1$. The term $B(\boldsymbol{x})$ doesn't depend on $\boldsymbol{\phi}$, and can usually be ignored. Can you see why we need this term at all?[8]

The notation $\text{p}(\boldsymbol{x} \mid y; \boldsymbol{\phi})$ indicates the conditional probability of word counts $\boldsymbol{x}$ given label $y$, with parameter $\boldsymbol{\phi}$, which is equal to $\text{p}_{\text{mult}}(\boldsymbol{x}; \boldsymbol{\phi}_y)$. By specifying the multinomial distribution, we describe the **multinomial naïve Bayes** classifier. Why "naïve"? Because the multinomial distribution treats each word token independently: the probability mass function factorizes across the counts.[9]

### 1.1.1 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**, $\boldsymbol{x}$, this model generates a *sequence* of **tokens**, $\boldsymbol{w} = (w_1, w_2, \ldots, w_M)$. The distinction between types and tokens is critical: $x_j \in \{0, 1, 2, \ldots, M\}$ is the count of word type $j$ in the vocabulary, e.g., the number of times

[7] You are encouraged to verify that the multinomial distribution is identical to the categorical distribution when $\sum_j x_j = 1$.

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

[9] You 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 your 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 classifier

**for** Document $i \in \{1, 2, \ldots, N\}$ **do**:
  Draw the label $y^{(i)} \sim \text{Categorical}(\boldsymbol{\mu})$;
  **for** Token $m \in \{1, 2, \ldots, M_i\}$ **do**:
    Draw the token $w_m^{(i)} \mid y^{(i)} \sim \text{Categorical}(\boldsymbol{\phi}_{y^{(i)}})$.

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 $\boldsymbol{w}$ is a product of categorical probabilities. Algorithm 2 makes a conditional independence assumption: each token $w_m^{(i)}$ is independent of all other tokens $w_n^{(i)}$, 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 the probability under multinomial Naïve Bayes. The constant of proportionality is the factor $B(\boldsymbol{x})$, which appears in the multinomial distribution. Because $B(\boldsymbol{x}) \geq 1$, the probability for a vector of counts $\boldsymbol{x}$ is at least as large as the probability for a list of words $\boldsymbol{w}$ 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 $B(\boldsymbol{x})$ is equal to the total number of possible word orderings for count vector $\boldsymbol{x}$.

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

### 1.1.2 Prediction

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

$$\hat{y} = \underset{y}{\text{argmax}} \log \text{p}(\boldsymbol{x}, y; \boldsymbol{\mu}, \boldsymbol{\phi}) \quad [1.16]$$

$$= \underset{y}{\text{argmax}} \log \text{p}(\boldsymbol{x} \mid y; \boldsymbol{\phi}) + \log \text{p}(y; \boldsymbol{\mu}) \quad [1.17]$$

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

$$\log \mathrm{p}(\boldsymbol{x} \mid y; \boldsymbol{\phi}) + \log \mathrm{p}(y; \boldsymbol{\mu}) = \log \left[ B(\boldsymbol{x}) \prod_{j=1}^{V} \boldsymbol{\phi}_{y,j}^{x_j} \right] + \log \mu_y \quad [1.18]$$

$$= \log B(\boldsymbol{x}) + \sum_{j=1}^{V} x_j \log \phi_{y,j} + \log \mu_y \quad [1.19]$$

$$= \log B(\boldsymbol{x}) + \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}, y), \quad [1.20]$$

where

$$\boldsymbol{\theta} = [\boldsymbol{\theta}^{(1)}; \boldsymbol{\theta}^{(2)}; \ldots; \boldsymbol{\theta}^{(K)}] \quad [1.21]$$

$$\boldsymbol{\theta}^{(y)} = [\log \phi_{y,1}, \log \phi_{y,2}, \ldots, \log \phi_{y,V}, \log \mu_y]^\top \quad [1.22]$$

The feature function $\boldsymbol{f}(\boldsymbol{x}, y)$ is a vector of $V$ word counts and an offset, padded by zeros for the labels not equal to $y$ (see Equations 1.3-1.6, and Figure 1.1). This construction ensures that the inner product $\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}, y)$ only activates the features whose weights are in $\boldsymbol{\theta}^{(y)}$. These features and weights are all we need to compute the joint log-probability $\log \mathrm{p}(\boldsymbol{x}, y)$ for each $y$. This is a key point: through this notation, we have converted the problem of computing the log-likelihood for a document-label pair $\langle \boldsymbol{x}, y \rangle$ into the computation of a vector inner product.

### 1.1.3 Estimation

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

$$\phi_{y,j} = \frac{\sum_{i:y^{(i)}=y} x_j^{(i)}}{\sum_{j'=1}^{V} \sum_{i:y^{(i)}=y} x_{j'}^{(i)}} = \frac{\text{count}(y, j)}{\sum_{j'=1}^{V} \text{count}(y, j')}, \quad [1.23]$$

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

Equation 1.23 defines the **relative frequency estimator** for $\boldsymbol{\phi}$. It can be justified as a **maximum likelihood estimate**: the estimate that maximizes the probability $\mathrm{p}(\boldsymbol{x}^{(1:N)}, y^{(1:N)}; \boldsymbol{\theta})$. Based on the generative model in Algorithm 1, the log-likelihood is,

$$\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\mu}) = \sum_{i=1}^{N} \log \mathrm{p}_{\text{mult}}(\boldsymbol{x}^{(i)}; \boldsymbol{\phi}_{y^{(i)}}) + \log \mathrm{p}_{\text{cat}}(y^{(i)}; \boldsymbol{\mu}), \quad [1.24]$$

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

$$\mathcal{L}(\boldsymbol{\phi}) = \sum_{i=1}^{N} \log \text{p}_{\text{mult}}(\boldsymbol{x}^{(i)}; \boldsymbol{\phi}_{y^{(i)}}) = \sum_{i=1}^{N} \log B(\boldsymbol{x}^{(i)}) + \sum_{j=1}^{V} x_j^{(i)} \log \phi_{y^{(i)},j}, \quad [1.25]$$

where $B(\boldsymbol{x}^{(i)})$ is constant with respect to $\boldsymbol{\phi}$.

We would now like to optimize the log-likelihood $\mathcal{L}$, by taking derivatives with respect to $\boldsymbol{\phi}$. But before we can do that, we have to deal with a set of constraints:

$$\sum_{j=1}^{V} \phi_{y,j} = 1 \quad \forall y \quad [1.26]$$

These constraints can be incorporated by adding a Lagrange multiplier. Solving separately for each label $y$, we obtain the Lagrangian,

$$\ell(\boldsymbol{\phi}_y) = \sum_{i:y^{(i)}=y} \sum_{j=1}^{V} x_j^{(i)} \log \phi_{y,j} - \lambda(\sum_{j=1}^{V} \phi_{y,j} - 1). \quad [1.27]$$

It is now possible to differentiate the Lagrangian with respect to the parameter of interest,

$$\frac{\partial \ell(\boldsymbol{\phi}_y)}{\partial \phi_{y,j}} = \sum_{i:y^{(i)}=y} x_j^{(i)} / \phi_{y,j} - \lambda \quad [1.28]$$

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

$$\lambda \phi_{y,j} = \sum_{i:y^{(i)}=y} x_j^{(i)} \quad [1.29]$$

$$\phi_{y,j} \propto \sum_{i:y^{(i)}=y} x_j^{(i)} = \sum_{i=1}^{N} \delta\left(y^{(i)} = y\right) x_j^{(i)} = \text{count}(y, j), \quad [1.30]$$

where $\delta\left(y^{(i)} = y\right)$ is a **delta function**, also sometimes called an **indicator function**, which returns one if $y^{(i)} = y$, and zero otherwise. Equation 1.30 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 $\boldsymbol{\phi}_y$ up to a constant of proportionality. Now recall the constraint $\sum_{j=1}^{V} \boldsymbol{\phi}_{y,j} = 1$, which arises because $\boldsymbol{\phi}_y$ represents a vector of probabilities for each word in the vocabulary. This constraint leads to an exact solution,

$$\phi_{y,j} = \frac{\text{count}(y, j)}{\sum_{j'=1}^{V} \text{count}(y, j')}. \quad [1.31]$$

This is equal to the relative frequency estimator from Equation 1.23. A similar derivation gives $\mu_y \propto \sum_{i=1}^{N} \delta\left(y^{(i)} = y\right)$.

### 1.1.4 Smoothing and MAP estimation

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 *Bayesian* may have never yet appeared in a spam email. But choosing a value of $\phi_{\text{SPAM},\textit{Bayesian}} = 0$ would allow this single feature to completely veto a label, since $\text{p}(\text{SPAM} \mid \boldsymbol{x}) = 0$ if $\boldsymbol{x}_{\textit{Bayesian}} > 0$.

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 $\alpha$ to each count, and then normalizing.

$$\phi_{y,j} = \frac{\alpha + \text{count}(y, j)}{V\alpha + \sum_{j'=1}^{V} \text{count}(y, j')} \quad [1.32]$$

This is called **Laplace smoothing**.[10] The pseudocount $\alpha$ is a **hyperparameter**, because it controls the form of the log-likelihood function, which in turn drives the estimation of $\boldsymbol{\phi}$.

Smoothing reduces variance, but it takes us away from the maximum likelihood estimate: it imposes a **bias**. In this case, the bias points towards uniform probabilities. Machine learning theory shows that errors on heldout data can be attributed to the sum of bias and variance (Mohri et al., 2012). Techniques for reducing variance typically increase the bias, leading to a **bias-variance tradeoff**.

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

### 1.1.5 Setting hyperparameters

How should we choose the best value of hyperparameters like $\alpha$? Maximum likelihood will not work: the maximum likelihood estimate of $\alpha$ on the training set will always be $\alpha = 0$. In many cases, what we really want is **accuracy**: the number of correct predictions, divided by the total number of predictions. (Other measures of classification performance are discussed in § 3.4.) As we will see, it is hard to optimize for accuracy directly. But for scalar hyperparameters like $\alpha$, we can employ a simple heuristic called **grid search**: try a

[10] Laplace smoothing has a nice Bayesian justification, in which the generative model is extended to include $\boldsymbol{\phi}$ as a random variable. The resulting estimate is called **maximum a posteriori**, or MAP.

set of values (e.g., $\alpha \in \{0.001, 0.01, 0.1, 1, 10\}$), compute the accuracy for each value, and choose the setting that maximizes the accuracy.

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

We will also want to predict the performance of our classifier on unseen data. To do this, we must hold out a separate subset of data, called the **test set**. It is critical that the test set not overlap with either the training or development sets, or else we risk overestimating the performance that the classifier will achieve on unlabeled data in the future. The test set should also not be used when making modeling decisions, such as the form of the feature function, the size of the vocabulary, and so on (these decisions are reviewed in chapter 3.) 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.

## 1.2 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** $\boldsymbol{f}(\boldsymbol{x}, y)$ 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, we make the following approximation:[11]

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

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

$$\begin{aligned} \Pr(\text{word} = \textit{unfit}, \text{prefix} = \textit{un-} \mid y) &= \Pr(\text{prefix} = \textit{un-} \mid \text{word} = \textit{unfit}, y) && [1.33] \\ &\quad \times \Pr(\text{word} = \textit{unfit} \mid y) && [1.34] \end{aligned}$$

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

$$\begin{aligned} \Pr(\text{word} = \textit{unfit}, \text{prefix} = \textit{un-} \mid y) &= 1 \qquad\qquad \times \Pr(\text{word} = \textit{unfit} \mid y) && [1.35] \\ &\gg \Pr(\text{prefix} = \textit{un-} \mid y) \quad \times \Pr(\text{word} = \textit{unfit} \mid y), && [1.36] \end{aligned}$$

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, $\text{p}(\boldsymbol{x}^{(1:N)}, y^{(1:N)})$, which requires modeling the probability of the observed text. In classification problems, we are always given $\boldsymbol{x}$, and are only interested in predicting the label $y$, so it seems unnecessary to model the probability of $\boldsymbol{x}$. **Discriminative learning** algorithms focus on the problem of predicting $y$, and do not attempt to model the probability of the text $\boldsymbol{x}$.

### 1.2.1 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

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

**Algorithm 3** Perceptron learning algorithm

```
procedure PERCEPTRON(x^(1:N), y^(1:N))
    t ← 0
    θ^(0) ← 0
    repeat
        t ← t + 1
        Select an instance i
        ŷ ← argmax_y θ^(t−1) · f(x^(i), y)
        if ŷ ≠ y^(i) then
            θ^(t) ← θ^(t−1) + f(x^(i), y^(i)) − f(x^(i), ŷ)
        else
            θ^(t) ← θ^(t−1)
    until tired
    return θ^(t)
```

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.

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

The perceptron algorithm may seem like a cheap heuristic: Naïve Bayes has a solid foundation in probability, but now we are 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**:

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

$$\exists \boldsymbol{\theta}, \rho > 0 : \forall \langle \boldsymbol{x}^{(i)}, y^{(i)} \rangle \in \mathcal{D}, \quad \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y^{(i)}) \geq \rho + \max_{y' \neq y^{(i)}} \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y'). \tag{1.37}$$

Linear separability is important because of the following guarantee: if your data is

linearly separable, then the perceptron algorithm will find a separator (Novikoff, 1962).[12] So while the perceptron may seem heuristic, it is guaranteed to succeed, if the learning problem is easy enough.

How useful is this proof? Minsky and Papert (1969) famously proved that the simple logical function of *exclusive-or* is not separable, and that a perceptron is therefore incapable of learning this function. But this is not just an issue for the perceptron: any linear classification algorithm, including Naïve Bayes, will fail on this task. In natural language classification problems usually involve high dimensional feature spaces, with thousands or millions of features. For these problems, it is very likely that the training data is indeed separable. And even if the data is not separable, it is still possible to place an upper bound on the number of errors that the perceptron algorithm will make (Freund and Schapire, 1999).

### 1.2.2 Averaged perceptron

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

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

Even if the data is not separable, the averaged weights will eventually converge. One possible stopping criterion is to check the difference between the average weight vectors after each pass through the data: if the norm of the difference falls below some predefined threshold, we can stop iterating. Another stopping criterion is to hold out some data, and to measure the predictive accuracy on this heldout data. When the accuracy on the heldout data starts to decrease, the learning algorithm has begun to **overfit** the training

[12] It 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 **statistical learning theory** (Mohri et al., 2012).

**Algorithm 4** Averaged perceptron learning algorithm

**procedure** AVG-PERCEPTRON($\boldsymbol{x}^{(1:N)}, \boldsymbol{y}^{(1:N)}$)
  $t \leftarrow 0$
  $\boldsymbol{\theta}^{(0)} \leftarrow 0$
  **repeat**
    $t \leftarrow t + 1$
    Select an instance $i$
    $\hat{y} \leftarrow \text{argmax}_y \boldsymbol{\theta}^{(t-1)} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y)$
    **if** $\hat{y} \neq y^{(i)}$ **then**
      $\boldsymbol{\theta}^{(t)} \leftarrow \boldsymbol{\theta}^{(t-1)} + \boldsymbol{f}(\boldsymbol{x}^{(i)}, y^{(i)}) - \boldsymbol{f}(\boldsymbol{x}^{(i)}, \hat{y})$
    **else**
      $\boldsymbol{\theta}^{(t)} \leftarrow \boldsymbol{\theta}^{(t-1)}$
    $\boldsymbol{m} \leftarrow \boldsymbol{m} + \boldsymbol{\theta}^{(t)}$
  **until** tired
  $\overline{\boldsymbol{\theta}} \leftarrow \frac{1}{t}\boldsymbol{m}$
  **return** $\overline{\boldsymbol{\theta}}$

set. At this point, it is probably best to stop; this stopping criterion is known as **early stopping**.

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

## 1.3 Loss functions and large-margin classification

Naïve Bayes chooses the weights $\boldsymbol{\theta}$ by maximizing the joint log-likelihood $\log \text{p}(\boldsymbol{x}^{(1:N)}, y^{(1:N)})$. Optimization problems are generally formulated as minimization of a **loss function**. We can trivially reformulate maximum likelihood in this way, by defining the loss function to be the *negative* log-likelihood:

$$\log \text{p}(\boldsymbol{x}^{(1:N)}, y^{(1:N)}; \boldsymbol{\theta}) = \sum_{i=1}^{N} \log \text{p}(\boldsymbol{x}^{(i)}, y^{(i)}; \boldsymbol{\theta}) \quad [1.38]$$

$$\ell_{\text{NB}}(\boldsymbol{\theta}; \boldsymbol{x}^{(i)}, y^{(i)}) = -\log \text{p}(\boldsymbol{x}^{(i)}, y^{(i)}; \boldsymbol{\theta}) \quad [1.39]$$

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{argmin}} \sum_{i=1}^{N} \ell_{\text{NB}}(\boldsymbol{\theta}; \boldsymbol{x}^{(i)}, y^{(i)}), \quad [1.40]$$

This minimization problem is identical to maximum-likelihood estimation. Loss functions provide a general framework for comparing machine learning objectives. For exam-

ple, an alternative loss function is the **zero-one loss**,

$$\ell_{\text{0-1}}(\boldsymbol{\theta}; \boldsymbol{x}^{(i)}, y^{(i)}) = \begin{cases} 0, & y^{(i)} = \text{argmax}_y \, \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y) \\ 1, & \text{otherwise} \end{cases} \quad [1.41]$$

This loss function is proportional to the error rate of the classifier on the training data. Since a low error rate is often the ultimate goal of classification, this may seem ideal. But the zero-one loss is **non-convex**,[13] which means that it is combinatorially difficult to optimize.

The perceptron optimizes the following loss function:

$$\ell_{\text{PERCEPTRON}}(\boldsymbol{\theta}; \boldsymbol{x}^{(i)}, y^{(i)}) = \max_{y \in \mathcal{Y}} \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y) - \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y^{(i)}), \quad [1.42]$$

which is a **hinge loss** with the hinge point at zero. When $\hat{y} = y^{(i)}$, the loss is zero; otherwise, it increases linearly with the gap between the score for the predicted label $\hat{y}$ and the score for the true label $y^{(i)}$.

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

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ell_{\text{PERCEPTRON}}(\boldsymbol{\theta}; \boldsymbol{x}^{(i)}, y^{(i)}) = \boldsymbol{f}(\boldsymbol{x}^{(i)}, \hat{y}) - \boldsymbol{f}(\boldsymbol{x}^{(i)}, y^{(i)}). \quad [1.43]$$

At each instance perceptron algorithm takes a step of magnitude one in the opposite direction of this **gradient**, $\nabla_{\boldsymbol{\theta}} \ell_{\text{PERCEPTRON}} = \frac{\partial}{\partial \boldsymbol{\theta}} \ell_{\text{PERCEPTRON}}(\boldsymbol{\theta}; \boldsymbol{x}^{(i)}, y^{(i)})$. As we will see in § 1.5, this is an example of the optimization algorithm **stochastic gradient descent**, applied to the objective in Equation 1.42.

**Breaking ties with subgradient descent** Careful readers will notice the tacit assumption that there is a unique $\hat{y}$ that maximizes $\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y)$. What if there are two or more labels that maximize this function? Consider binary classification: if the maximizer is $y^{(i)}$, then the gradient is zero, and so is the perceptron update; if the maximizer is $\hat{y} \neq y^{(i)}$, then the update is the difference $\boldsymbol{f}(\boldsymbol{x}^{(i)}, y^{(i)}) - \boldsymbol{f}(\boldsymbol{x}^{(i)}, \hat{y})$. The underlying issue is that the perceptron loss is not **smooth**, because the first derivative has a discontinuity at the hinge point, where the score for the true label $y^{(i)}$ is equal to the score for some other label $\hat{y}$. At this point, there is no unique gradient; rather, there is a set of **subgradients**. A vector $\boldsymbol{v}$ is a subgradient of the function $g$ at $\boldsymbol{u}_0$ iff $g(\boldsymbol{u}) - g(\boldsymbol{u}_0) \geq \boldsymbol{v} \cdot (\boldsymbol{u} - \boldsymbol{u}_0)$ for all $\boldsymbol{u}$. Graphically, this defines the set of hyperplanes that include $g(\boldsymbol{u}_0)$ and do not intersect $g$ at any other point. As we approach the hinge point from the left, the gradient is $\boldsymbol{f}(\boldsymbol{x}, \hat{y}) - \boldsymbol{f}(\boldsymbol{x}, y)$; as we

[13] A 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. 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).

approach from the right, the gradient is $\mathbf{0}$. At the hinge point, the subgradients include all vectors that are bounded by these two extremes. In subgradient descent, *any* subgradient can be used (Bertsekas, 2012). Since both $\mathbf{0}$ and $\boldsymbol{f}(\boldsymbol{x}, \hat{y}) - \boldsymbol{f}(\boldsymbol{x}, y)$ are subgradients at the hinge point, either one can be used in the perceptron update.

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

- Both $\ell_{\text{NB}}$ and $\ell_{\text{PERCEPTRON}}$ are convex, making them relatively easy to optimize. However, $\ell_{\text{NB}}$ can be optimized in closed form, while $\ell_{\text{PERCEPTRON}}$ requires iterating over the dataset multiple times.
- $\ell_{\text{NB}}$ can suffer **infinite** loss on a single example, since the logarithm of zero probability is negative infinity. Naïve Bayes will therefore overemphasize some examples, and underemphasize others.
- $\ell_{\text{PERCEPTRON}}$ treats all correct answers equally. Even if $\boldsymbol{\theta}$ only gives the correct answer by a tiny margin, the loss is still zero.

### 1.3.1 Large margin classification

This last comment suggests a potential problem with the perceptron. Suppose a test example is very close to a training example, but not identical. If the classifier only gets the correct answer on the training example by a small margin, then it may get the test instance wrong. To formalize this intuition, define the **margin** as,

$$\gamma(\boldsymbol{\theta}; \boldsymbol{x}^{(i)}, y^{(i)}) = \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y^{(i)}) - \max_{y \neq y^{(i)}} \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y). \quad [1.44]$$

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

$$\ell_{\text{MARGIN}}(\boldsymbol{\theta}; \boldsymbol{x}^{(i)}, y^{(i)}) = \begin{cases} 0, & \gamma(\boldsymbol{\theta}; \boldsymbol{x}^{(i)}, y^{(i)}) \geq 1, \\ 1 - \gamma(\boldsymbol{\theta}; \boldsymbol{x}^{(i)}, y^{(i)}), & \text{otherwise} \end{cases} \quad [1.45]$$

$$= \Big(1 - \gamma(\boldsymbol{\theta}; \boldsymbol{x}^{(i)}, y^{(i)})\Big)_+, \quad [1.46]$$

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 1.2. The margin loss is a convex upper bound on the zero-one loss.

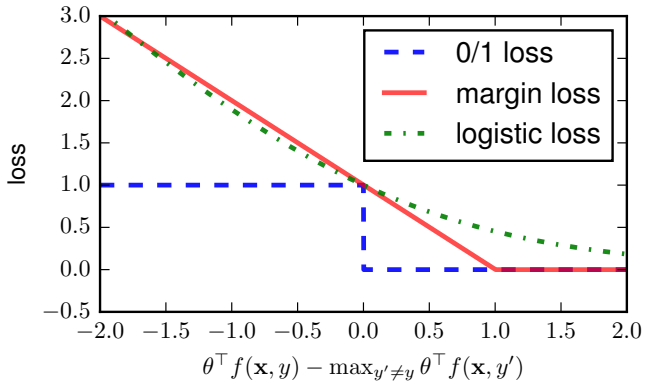



Figure 1.2: Margin, zero-one, and logistic loss functions.[todo: redo x-axis labels, add hinge loss with hinge point at zero.]

### 1.3.2 Support vector machines

If a dataset is linearly separable, then there is some hyperplane $\boldsymbol{\theta}$ that correctly classifies all training instances with margin $\rho$ (by Definition 1). This margin can be increased to any desired value by multiplying the weights by a constant. Now, for any datapoint $\langle \boldsymbol{x}^{(i)}, y^{(i)} \rangle$, the geometric distance to the separating hyperplane is given by $\frac{\gamma(\boldsymbol{\theta}; \boldsymbol{x}^{(i)}, y^{(i)})}{||\boldsymbol{\theta}||_2}$, where the denominator is the norm of the weights, $||\boldsymbol{\theta}||_2 = \sqrt{\sum_j \theta_j^2}$. The geometric distance is sometimes called the **geometric margin**, in contrast to the **functional margin** $\gamma(\boldsymbol{\theta}; \boldsymbol{x}^{(i)}, y^{(i)})$. Both are shown in Figure 1.3. The geometric margin is a good measure of the robustness of the separator: if the functional margin is large, but the norm $||\boldsymbol{\theta}||_2$ is also large, then a small change in $\boldsymbol{x}^{(i)}$ could cause it to be misclassified. We therefore seek to maximize the minimum geometric margin, subject to the constraint that the functional margin is at least one:

$$\begin{aligned} \max_{\boldsymbol{\theta}} . \quad & \min_i . \quad \frac{\gamma(\boldsymbol{\theta}; \boldsymbol{x}^{(i)}, y^{(i)})}{||\boldsymbol{\theta}||_2} \\ \text{s.t.} \quad & \gamma(\boldsymbol{\theta}; \boldsymbol{x}^{(i)}, y^{(i)}) \geq 1, \quad \forall i. \end{aligned} \tag{1.47}$$

This is a **constrained optimization** problem, where the second line describes constraints on the space of possible solutions $\boldsymbol{\theta}$. In this case, the constraint is that the functional margin always be at least one, and the objective is that the minimum geometric margin be as large as possible.

Any scaling factor on $\boldsymbol{\theta}$ will cancel in the numerator and denominator of the geometric margin. This means that if the data is linearly separable at $\rho$, we can increase this margin

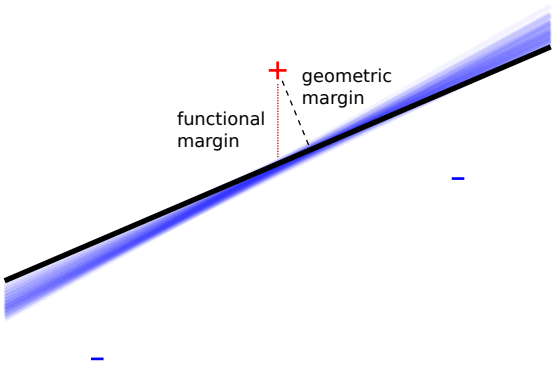



Figure 1.3: Functional and geometric margins for a binary classification problem. All separators that satisfy the margin constraint are shown. The separator with the largest geometric margin is shown in bold.

to 1 by rescaling $\boldsymbol{\theta}$. We therefore need only minimize the denominator $||\boldsymbol{\theta}||_2$, subject to the constraint on the functional margin. The minimizer of $||\boldsymbol{\theta}||_2$ is also the minimizer of $\frac{1}{2}||\boldsymbol{\theta}||_2^2 = \frac{1}{2}\sum_j \theta_J^2$, which is easier to work with. This gives the optimization problem,

$$\begin{aligned} \min_{\boldsymbol{\theta}} . \quad & \frac{1}{2}||\boldsymbol{\theta}||_2^2 \\ \text{s.t.} \quad & \gamma(\boldsymbol{\theta}; \boldsymbol{x}^{(i)}, y^{(i)}) \geq 1, \quad \forall_i. \end{aligned} \tag{1.48}$$

This optimization problem is a **quadratic program**: the objective is a quadratic function of the parameters, and the constraints are all linear inequalities. The resulting classifier is better known as the **support vector machine**. The name derives from one of the solutions, which is to incorporate the constraints through Lagrange variables $\alpha_i \geq 0, i \in \{1, 2, \ldots, N\}$. The instances for which $\alpha_i > 0$ are the **support vectors**; other instances are irrelevant to the classification boundary.

### 1.3.3 Slack variables

If a dataset is not linearly separable, then there is no $\boldsymbol{\theta}$ that satisfies the margin constraint. To add more flexibility, we 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:

$$\begin{aligned} \min_{\boldsymbol{\theta}, \boldsymbol{\xi}} \quad & \frac{1}{2}||\boldsymbol{\theta}||_2^2 + C \sum_{i=1}^{N} \xi_i \\ \text{s.t.} \quad & \gamma(\boldsymbol{\theta}; \boldsymbol{x}^{(i)}, y^{(i)}) + \xi_i \geq 1, \quad \forall i \\ & \xi_i \geq 0, \quad \forall_i. \end{aligned} \tag{1.49}$$

The hyperparameter $C$ controls the tradeoff between violations of the margin constraint and the preference for a low norm of $\boldsymbol{\theta}$. As $C \rightarrow \infty$, slack is infinitely expensive, and there is only a solution if the data is separable. As $C \rightarrow 0$, slack becomes free, and there is a trivial solution at $\boldsymbol{\theta} = \mathbf{0}$. Thus, $C$ plays a similar role to the smoothing parameter in Naïve Bayes (§ 1.1.4), trading off between a close fit to the training data and better generalization. Like the smoothing parameter of Naïve Bayes, $C$ must be set by the user, typically by maximizing performance on a heldout development set.

To solve the constrained optimization problem defined in Equation 1.49, we can first solve for the slack variables,

$$\xi_i \geq (1 - \gamma(\boldsymbol{\theta}; \boldsymbol{x}^{(i)}, y^{(i)}))_+. \tag{1.50}$$

The inequality is tight, because the slack variables are penalized in the objective, and there is no advantage to increasing them beyond the minimum value (Ratliff et al., 2007; Smith, 2011). The problem can therefore be transformed into the unconstrained optimization,

$$\min_{\boldsymbol{\theta}} \quad \frac{\lambda}{2}||\boldsymbol{\theta}||_2^2 + \sum_{i=1}^{N} (1 - \gamma(\boldsymbol{\theta}; \boldsymbol{x}^{(i)}, y^{(i)}))_+, \tag{1.51}$$

where each $\xi_i$ has been substituted by the right-hand side of Equation 1.50, and the factor of $C$ on the slack variables has been replaced by an equivalent factor of $\lambda = \frac{1}{C}$ on the norm of the weights.

Now define the **cost** of a classification error as,[14]

$$c(y^{(i)}, \hat{y}) = \begin{cases} 1, & y^{(i)} \neq \hat{y} \\ 0, & \text{otherwise.} \end{cases} \tag{1.52}$$

We can then rewrite Equation 1.51,

$$\min_{\boldsymbol{\theta}} \quad \frac{\lambda}{2}||\boldsymbol{\theta}||_2^2 + \sum_{i=1}^{N} \left( \max_{y \in \mathcal{Y}} \Big( \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y) + c(y^{(i)}, y) \Big) - \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y^{(i)}) \right)_+, \tag{1.53}$$

[14] We can also define specialized cost functions that heavily penalize especially undesirable errors (Tsochantaridis et al., 2004). This issue is revisited in chapter 6.

maximizing over all $y \in \mathcal{Y}$, in search of labels that are both *high scoring*, as measured by $\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y)$, and *wrong*, as measured by $c(y^{(i)}, y)$. This maximization is known as **cost-augmented inference**, because it augments the maximization objective to favor high-cost predictions. If the highest-scoring label is $y = y^{(i)}$, then the margin constraint is satisfied, and the loss for this instance is zero. Note that cost-augmentation is not applied when making predictions on unseen data.

Differentiating Equation 1.53 with respect to the weights gives,

$$\nabla_{\boldsymbol{\theta}} L_{\text{SVM}} = \lambda \boldsymbol{\theta} + \sum_{i=1}^{N} \boldsymbol{f}(\boldsymbol{x}^{(i)}, \hat{y}) - \boldsymbol{f}(\boldsymbol{x}^{(i)}, y^{(i)}) \quad [1.54]$$

$$\hat{y} = \underset{y \in \mathcal{Y}}{\text{argmax}}\, \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y) + c(y^{(i)}, y), \quad [1.55]$$

where $L_{\text{SVM}}$ refers to minimization objective in Equation 1.53. This gradient is very similar to the perceptron update. One difference is the additional term $\lambda\boldsymbol{\theta}$, which **regularizes** the weights towards $\mathbf{0}$. The other difference is the cost $c(y^{(i)}, y)$, which is added to $\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}, y)$ when choosing $\hat{y}$ during training. This term derives from the margin constraint: large margin classifiers learn not only from instances that are incorrectly classified, but also from instances for which the correct classification decision was not sufficiently confident.

## 1.4 Logistic regression

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

**Logistic regression** combines advantages of discriminative and probabilistic classifiers. Unlike Naïve Bayes, which starts from the **joint probability** $\text{p}(\boldsymbol{x}, y)$, logistic regression defines the desired **conditional probability** $\text{p}(y \mid \boldsymbol{x})$ directly. Think of $\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}, y)$ as a scoring function for the compatibility of the base features $\boldsymbol{x}$ and the label $y$. To convert this score into a probability, we first exponentiate, obtaining $\exp(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}, y))$, 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,

$$\text{p}(y \mid \boldsymbol{x}; \boldsymbol{\theta}) = \frac{\exp(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}, y))}{\sum_{y' \in \mathcal{Y}} \exp(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}, y'))}. \quad [1.56]$$

Given a dataset $\mathcal{D} = \{\boldsymbol{x}^{(1:N)}, y^{(1:N)}\}$, the weights $\boldsymbol{\theta}$ are estimated by **maximum conditional likelihood**,

$$\log \mathrm{p}(\boldsymbol{y}^{(1:N)} \mid \boldsymbol{x}^{(1:N)}; \boldsymbol{\theta}) = \sum_{i=1}^{N} \log \mathrm{p}(y^{(i)} \mid \boldsymbol{x}^{(i)}; \boldsymbol{\theta}) \quad [1.57]$$

$$= \sum_{i=1}^{N} \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y^{(i)}) - \log \sum_{y' \in \mathcal{Y}} \exp\left(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y')\right). \quad [1.58]$$

The final line is obtained by plugging in Equation 1.56 and taking the logarithm.[15] Inside the sum, we have the (additive inverse of the) **logistic loss**,

$$\ell_{\text{LOGREG}}(\boldsymbol{\theta}; \boldsymbol{x}^{(i)}, y^{(i)}) = -(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y^{(i)}) - \log \sum_{y' \in \mathcal{Y}} \exp(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y'))) \quad [1.59]$$

The logistic loss is shown in Figure 1.2. A key difference from the zero-one and hinge losses is that logistic loss is never exactly zero. This means that the objective function can always be improved by choosing the correct label with more confidence.

### 1.4.1 Regularization

As with the margin-based algorithms described in § 1.3, we can obtain better generalization by penalizing the norm of $\boldsymbol{\theta}$. This is done by adding a term of $\frac{\lambda}{2}||\boldsymbol{\theta}||_2^2$ to the minimization objective. This is called $L_2$ regularization, because $||\boldsymbol{\theta}||_2^2$ is the squared $L_2$ norm of the vector $\boldsymbol{\theta}$. $L_2$ regularization can be viewed as placing a zero-mean Gaussian prior distribution on each term of $\boldsymbol{\theta}$, because the log-likelihood under a zero-mean Gaussian is,

$$\log N(\theta_j; 0, \sigma^2) \propto -\frac{1}{2\sigma^2}\theta_j^2, \quad [1.60]$$

so that $\lambda = \frac{1}{\sigma^2}$.

As in the support vector machine classifier, $L_2$ regularization forces the estimator to trade performance on the training data for a smaller norm of the weights, and this can help to prevent overfitting. Regularization is generally considered to be essential to estimating high-dimensional models. Consider what would happen to the unregularized weight for a base feature $j$ that was active in only one instance $\boldsymbol{x}^{(i)}$: the conditional likelihood could always be improved by increasing the weight for this feature, so that $\boldsymbol{\theta}_{(j,y^{(i)})} \rightarrow \infty$ and $\boldsymbol{\theta}_{(j,\tilde{y} \neq y^{(i)})} \rightarrow -\infty$, where $(j, y)$ is the index of feature associated with $x_j^{(i)}$ and label $y$ in $\boldsymbol{f}(\boldsymbol{x}^{(i)}, y)$.

[15] The 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.

### 1.4.2 Gradients

For a single example, here is the logistic loss and its gradient:

$$\ell_{\text{LOGREG}} = -\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y^{(i)}) + \log \sum_{y' \in \mathcal{Y}} \exp\Big(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y')\Big) \tag{1.61}$$

$$\frac{\partial \ell}{\partial \boldsymbol{\theta}} = -\boldsymbol{f}(\boldsymbol{x}^{(i)}, y^{(i)}) + \frac{1}{\sum_{y'' \in \mathcal{Y}} \exp\big(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y'')\big)} \times \sum_{y' \in \mathcal{Y}} \exp\Big(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y')\Big) \times \boldsymbol{f}(\boldsymbol{x}^{(i)}, y') \tag{1.62}$$

$$= -\boldsymbol{f}(\boldsymbol{x}^{(i)}, y^{(i)}) + \sum_{y' \in \mathcal{Y}} \frac{\exp\big(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y')\big)}{\sum_{y'' \in \mathcal{Y}} \exp\big(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y'')\big)} \times \boldsymbol{f}(\boldsymbol{x}^{(i)}, y') \tag{1.63}$$

$$= -\boldsymbol{f}(\boldsymbol{x}^{(i)}, y^{(i)}) + \sum_{y' \in \mathcal{Y}} \text{p}(y' \mid \boldsymbol{x}^{(i)}; \boldsymbol{\theta}) \times \boldsymbol{f}(\boldsymbol{x}^{(i)}, y') \tag{1.64}$$

$$= -\boldsymbol{f}(\boldsymbol{x}^{(i)}, y^{(i)}) + E_{y|\boldsymbol{x}^{(i)}}[\boldsymbol{f}(\boldsymbol{x}^{(i)}, y)]. \tag{1.65}$$

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

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

$$L_{\text{LOGREG}} = \frac{\lambda}{2}||\boldsymbol{\theta}||_2^2 - \sum_{i=1}^{N} \left(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y^{(i)}) - \log \sum_{y' \in \mathcal{Y}} \exp \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y')\right) \tag{1.66}$$

$$\nabla_{\boldsymbol{\theta}} L_{\text{LOGREG}} = \lambda\boldsymbol{\theta} - \sum_{i=1}^{N} \Big(\boldsymbol{f}(\boldsymbol{x}^{(i)}, y^{(i)}) - E_{y|\boldsymbol{x}}[\boldsymbol{f}(\boldsymbol{x}^{(i)}, y)]\Big). \tag{1.67}$$

## 1.5 Optimization

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

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

- In the support vector machine, the objective is,

$$L_{\text{SVM}} = \frac{\lambda}{2}||\boldsymbol{\theta}||_2^2 + \sum_{i=1}^{N}(\max_{y\in\mathcal{Y}}(\boldsymbol{\theta}\cdot\boldsymbol{f}(\boldsymbol{x}^{(i)},y) + c(y^{(i)},y)) - \boldsymbol{\theta}\cdot\boldsymbol{f}(\boldsymbol{x}^{(i)},y^{(i)}))_+, \quad [1.68]$$

  There is no closed-form solution, but the objective is convex. The perceptron algorithm minimizes a similar objective.

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

$$L_{\text{LOGREG}} = \frac{\lambda}{2}||\boldsymbol{\theta}||_2^2 - \sum_{i=1}^{N}\left(\boldsymbol{\theta}\cdot\boldsymbol{f}(\boldsymbol{x}^{(i)},y^{(i)}) - \log\sum_{y\in\mathcal{Y}}\exp\left(\boldsymbol{\theta}\cdot\boldsymbol{f}(\boldsymbol{x}^{(i)},y)\right)\right) \quad [1.69]$$

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

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

### 1.5.1 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,

$$\boldsymbol{\theta}^{(t+1)} \leftarrow \boldsymbol{\theta}^{(t)} - \eta^{(t)}\nabla_{\boldsymbol{\theta}}L \quad [1.70]$$

where $\nabla_{\boldsymbol{\theta}}L$ is the gradient computed over the entire training set, and $\eta^{(t)}$ is the **step size** at iteration $t$. If the objective $L$ is a convex function of $\boldsymbol{\theta}$, then this procedure is guaranteed to terminate at the global optimum, as long as $\eta^{(t)}$ is small enough.

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

$$H_{i,j} = \frac{\partial^2\ell}{\partial\theta_i\partial\theta_j} \quad [1.71]$$

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

For any gradient-based technique, the user must set the learning rates $\eta^{(t)}$. Convergence proofs often use the schedule $\eta^{(t)} = \eta^{(0)} t^{-\alpha}$ for $\alpha \in [1, 2]$. In practice, $\eta^{(t)}$ may also be fixed to a small constant, like $10^{-3}$. In either case, it is typical to try a few different values, and see which minimizes the objective most quickly.

### 1.5.2 Online optimization

Batch optimization computes the objective on the entire training set before making an update. This may be inefficient, because in many cases, a small number of training examples could point the learner in roughly the correct direction. **Online learning** algorithms make updates to the weights while iterating through the training data. The theoretical basis for this approach is a stochastic approximation to the true objective function,

$$\sum_{i=1}^{N} \ell(\boldsymbol{\theta}; \boldsymbol{x}^{(i)}, y^{(i)}) \approx N \times \ell(\boldsymbol{\theta}; \boldsymbol{x}^{(j)}, y^{(j)}), \qquad \langle \boldsymbol{x}^{(j)}, y^{(j)} \rangle \sim \{\boldsymbol{x}^{(1:N)}, y^{(1:N)}\}, \quad [1.72]$$

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

In **stochastic gradient descent**, we randomly sample a single instances, and make an immediate update based on the local gradient. Note the similarity to the perceptron algorithm, which also updates the weights one instance at a time. In **minibatch** stochastic gradient descent, the gradient is computed over a small set of instances. A typical approach is to set the minibatch size so that the entire batch fits in memory on the GPU (Neubig et al., 2017). It is then possible to speed up learning by parallelizing the computation of the gradient over each instance in the minibatch.

Algorithm 5 offers a generalized view of gradient descent. If the batcher returns a single batch with all instances, then the algorithm is standard gradient descent; if it returns $N$ batches with one instance each, then it is stochastic gradient descent; otherwise, the batcher can return $B$ minibatches, $1 < B < N$.

There are a number of more advanced techniques for online learning (Bottou et al., 2016). Many algorithms use an adaptive step size, 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 step size; rare features will be updated infrequently, so it is better to take larger steps. The **AdaGrad** (adaptive gradient) algorithm achieves this behavior by

**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, $1 < B < N$.

**procedure** GRADIENT-DESCENT($\boldsymbol{x}^{(1:N)}, \boldsymbol{y}^{(1:N)}, L, \eta^{(1...\infty)}$, Batcher, $T_{\text{max}}$)
  $\boldsymbol{\theta} \leftarrow \mathbf{0}$
  $t \leftarrow 0$
  **repeat**
    $\langle \boldsymbol{b}^{(1)}, \boldsymbol{b}^{(2)}, \ldots, \boldsymbol{b}^{(B)} \rangle \leftarrow \text{Batcher}(N)$
    **for** $n \in \{1, 2, \ldots, B\}$ **do**
      $t \leftarrow t + 1$
      $\boldsymbol{\theta}^{(t)} \leftarrow \boldsymbol{\theta}^{(t-1)} - \eta^{(t)} \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}^{(t-1)}; \boldsymbol{x}^{(b_1^{(n)}, b_2^{(n)}, \ldots)}, \boldsymbol{y}^{(b_1^{(n)}, b_2^{(n)}, \ldots)})$
      **if** Converged($\boldsymbol{\theta}^{(1,2,\ldots,t)}$) **then**
        **return** $\boldsymbol{\theta}^{(t)}$
  **until** $t \geq T_{\text{max}}$
  **return** $\boldsymbol{\theta}^{(t)}$

storing the sum of the squares of the gradients for each feature, and rescaling the learning rate by its inverse:

$$\boldsymbol{g}_t = \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}^{(t)}; \boldsymbol{x}^{(i)}, y^{(i)}) \tag{1.73}$$

$$\theta_j^{(t+1)} \leftarrow \theta_j^{(t)} - \frac{\eta^{(t)}}{\sqrt{\sum_{t'=1}^{t} g_{t,j}^2}} g_{t,j}, \tag{1.74}$$

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

In most cases, the number of active features for any instance is much smaller than the number of weights, $|\boldsymbol{f}(\boldsymbol{x}^{(i)}, y^{(i)})| \ll |\boldsymbol{\theta}|$. If so, the computation cost of online optimization will be dominated by the update from the regularization term, $\lambda\boldsymbol{\theta}$. The solution is to be "lazy", updating each $\theta_j$ only as it is used. To implement lazy updating, store an additional parameter $\tau_j$, which is the iteration at which $\theta_j$ was last updated. If $\theta_j$ is needed at time $t$, the $t - \tau$ regularization updates can be performed all at once. This strategy is described by Kummerfeld et al. (2015).

## 1.6 *Additional topics in classification

### 1.6.1 Sparsity-inducing regularization

In logistic regression and large-margin classification, generalization can be improved by regularizing the weights towards 0. The most popular regularizer is the squared $L_2$ norm

$||\boldsymbol{\theta}||_2^2$. However, we might prefer some other norm, such as $L_0 = ||\boldsymbol{\theta}||_0 = \sum_j \delta(\theta_j \neq 0)$, which applies a constant penalty for each non-zero weight. This norm can be thought of as a form of **feature selection**: optimizing the $L_0$-regularized conditional likelihood is equivalent to trading off the log-likelihood against the number of active features. Reducing the number of active features is desirable because the resulting model will be fast, low-memory, and should generalize well, since irrelevant features will be pruned away. Unfortunately, the $L_0$ norm is non-convex and non-differentiable; optimization under $L_0$ regularization is **NP-hard**, meaning that it can be solved efficiently only if P=NP (Ge et al., 2011).

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

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

### 1.6.2 Other views of logistic regression

In binary classification, we can dispense with the feature function, and choose $y$ based on the inner product of $\boldsymbol{\theta} \cdot \boldsymbol{x}$. We can obtain a conditional probability $\text{p}(y \mid \boldsymbol{x}; \boldsymbol{\theta})$ by passing this inner product through a **logistic function**,

$$\sigma(a) \triangleq \frac{\exp(a)}{1 + \exp(a)} = (1 + \exp(-a))^{-1} \quad [1.75]$$

$$\text{p}(y \mid \boldsymbol{x}; \boldsymbol{\theta}) = \sigma(\boldsymbol{\theta} \cdot \boldsymbol{x}). \quad [1.76]$$

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

In the early NLP literature, logistic regression is frequently called **maximum entropy** classification (Berger et al., 1996). This name can be traced to an alternative formulation, which tries to find the maximum entropy probability function that satisfies **moment-matching** constraints. These constraints specify that the empirical counts of each label-feature pair should match the expected counts under the induced probability distribution

$\mathrm{p}(y \mid \boldsymbol{x}^{(i)}; \boldsymbol{\theta})$,

$$\sum_{i=1}^{N} f_j(\boldsymbol{x}^{(i)}, y^{(i)}) = \sum_{i=1}^{N} \sum_{y \in \mathcal{Y}} \mathrm{p}(y \mid \boldsymbol{x}^{(i)}; \boldsymbol{\theta}) f_j(\boldsymbol{x}^{(i)}, y), \quad \forall j \tag{1.77}$$

This constraint will be met exactly when the derivative of the conditional log-likelihood function (Equation 1.65) is equal to zero. However, the constraint can be met by many values of $\boldsymbol{\theta}$, so which should we choose?

Let $\mathrm{p}_{y|\boldsymbol{x}}$ refer to the conditional distribution *function* of $y$ given $\boldsymbol{x}$, and $\mathrm{p}(y \mid \boldsymbol{x})$ to refer to the probability of a specific $y$ given a specific $\boldsymbol{x}$. The **entropy** of the conditional probability distribution $\mathrm{p}_{y|\boldsymbol{x}}$ is,

$$H(\mathrm{p}_{y|\boldsymbol{x}}) = -\sum_{\boldsymbol{x} \in \mathcal{X}} \mathrm{p}_{\boldsymbol{x}}(\boldsymbol{x}) \sum_{y \in \mathcal{Y}} \mathrm{p}_{y|\boldsymbol{x}}(y \mid \boldsymbol{x}) \log \mathrm{p}_{y|\boldsymbol{x}}(y \mid \boldsymbol{x}), \tag{1.78}$$

where $\mathcal{X}$ is the set of all possible feature vectors, and $\mathrm{p}_{\boldsymbol{x}}(\boldsymbol{x})$ is the probability of observing the base features $\boldsymbol{x}$. The distribution $\mathrm{p}_{\boldsymbol{x}}$ is unknown, but an empirical estimate can be computed by summing over all the instances in the training set,

$$\tilde{H}(\mathrm{p}_{y|\boldsymbol{x}}) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{y \in \mathcal{Y}} \mathrm{p}_{y|\boldsymbol{x}}(y \mid \boldsymbol{x}^{(i)}) \log \mathrm{p}_{y|\boldsymbol{x}}(y \mid \boldsymbol{x}^{(i)}). \tag{1.79}$$

If the entropy is large, the likelihood function is smooth across possible values of $y$; if it is small, the likelihood function is sharply peaked at some preferred value; in the limiting case, the entropy is zero if $\mathrm{p}(y \mid x) = 1$ for some $y$. The maximum-entropy criterion chooses to make the weakest commitments possible, while satisfying the moment-matching constraints from Equation 1.77. The solution to this constrained optimization problem is identical to the maximum conditional likelihood (logistic-loss) formulation we considered in the previous section.

### 1.6.3 Further reading

For more on stochastic gradient descent, as applied to a number of different learning algorithms, see (Zhang, 2004) and (Bottou, 1998). Murphy (2012) traces stochastic gradient descent to Nemirovski and Yudin (1978). Kummerfeld et al. (2015) empirically reviews several optimization algorithms for large-margin learning.

## 1.7 Summary of learning algorithms

Having seen several learning algorithms, it is natural to ask which is best in various situations.

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

**Perceptron** *Pros*: easy to implement; online; error-driven learning means that accuracy is typically high, especially after averaging. *Cons*: not probabilistic; non-averaged perceptron performs poorly if data is not separable; hard to know when to stop learning; lack of margin can lead to overfitting.

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

**Logistic regression** *Pros*: error-driven and probabilistic; overfitting is controlled by a regularization parameter. *Cons*: batch learning requires black-box optimization; logistic loss sometimes gives lower accuracy than hinge loss, due to overtraining on correctly-labeled examples.

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

## Exercises

1. Derive the maximum-likelihood estimate for the parameter $\boldsymbol{\mu}$ in Naïve Bayes.

2. As noted in the discussion of averaged perceptron in § 1.2.2, the computation of the running sum $\boldsymbol{m} \leftarrow \boldsymbol{m} + \boldsymbol{\theta}$ is unnecessarily expensive, requiring $K \times V$ operations. Give an alternative way to compute the averaged weights $\overline{\boldsymbol{\theta}}$, with complexity that is independent of $V$ and linear in the sum of feature sizes $\sum_{i=1}^{N} |\boldsymbol{f}(\boldsymbol{x}^{(i)}, y^{(i)})|$.

3. Consider a dataset that is comprised of two identical instances $\boldsymbol{x}^{(1)} = \boldsymbol{x}^{(2)}$ with distinct labels $y^{(1)} \neq y^{(2)}$. Assume all features are binary $x_j \in \{0, 1\}$ for all $j$.

   Now suppose that the averaged perceptron always chooses $i = 1$ when $t$ is even, and $i = 2$ when $t$ is odd, and that it will terminate under the following condition:

   $$\epsilon \geq \max_j \left| \frac{1}{t} \sum_t \theta_j^{(t)} - \frac{1}{t-1} \sum_t \theta_j^{(t-1)} \right|. \quad [1.80]$$

In words, the algorithm stops when the largest change in the averaged weights is less than or equal to $\epsilon$. Compute the number of iterations before the averaged perceptron terminates.

4. Suppose you have two labeled datasets $D_1$ and $D_2$, with the same features and labels.

   - Let $\boldsymbol{\theta}^{(1)}$ be the unregularized logistic regression (LR) coefficients from training on dataset $D_1$.
   - Let $\boldsymbol{\theta}^{(2)}$ be the unregularized LR coefficients (same model) from training on dataset $D_2$.
   - Let $\boldsymbol{\theta}^{*}$ be the unregularized LR coefficients from training on the combined dataset $D_1 \cup D_2$.

   Under these conditions, prove that for any feature $j$,

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

# Chapter 2

# Nonlinear classification

Linear classification may seem like all we need for natural language processing. The bag-of-words representation is inherently high-dimensional, and the number of features is often larger than the number of training instances. This means that we can usually find a linear classifier that perfectly fits the training data; moving to nonlinear classification may seem to only increase the risk of overfitting. For many tasks, lexical features are meaningful in isolation, and can offer independent evidence about the instance label — unlike computer vision, where individual pixels are rarely informative, and must be evaluated holistically to make sense of an image. For these reasons, natural language processing has historically focused on linear classification to a much greater extent than other machine learning application domains.

Yet in recent years, nonlinear classifiers have swept through natural language processing, and are now default approach for many tasks (Manning, 2016). There are at least three reasons for this change.

- There have been rapid advances in **deep learning**, a family of nonlinear methods that learn complex functions of the input through multiple layers of computation (Goodfellow et al., 2016).
- Deep learning facilitates the incorporation of **word embeddings**, which are dense vector representations of words. Word embeddings can be learned from large amounts of unlabeled data, and enable generalization to words that do not appear in the annotated training data (word embeddings are discussed in detail in chapter 13). This makes it possible to train high-coverage NLP systems with relatively little labeled training data.
- A third reason for the rise of deep nonlinear learners is hardware. Standard CPUs are no longer increasing in speed, but many deep learning models can be implemented efficiently on graphics processing units (GPUs), offering substantial perfor-

mance improvements.

This chapter focuses on **neural networks**, which are by far the dominant approach for nonlinear classification in natural language processing today.[1] Historically, a few other nonlinear learning methods have been applied to language data:

- **Kernel methods** are generalizations of the **nearest-neighbor** classification rule, which classifies each instance by the label of the most similar example in the training set (Hastie et al., 2009). The application of the **kernel support vector machine** to information extraction is described in chapter 16.
- **Decision trees** classify instances by checking a set of conditions. Scaling decision trees to bag-of-words inputs is difficult, but they have been effectively applied to problems such as coreference resolution (chapter 14), where more compact feature sets can be constructed (Soon et al., 2001).
- **Boosting** and related **ensemble methods** work by combining the predictions of several "weak" classifiers, each of which may consider only a small subset of features. Boosting has been successfully applied to text classification (Schapire and Singer, 2000) and syntactic analysis (Abney et al., 1999), and remains one of the most competitive methods for machine learning competition sites such as Kaggle (Chen and Guestrin, 2016).

## 2.1 Feedforward neural networks

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

1. **Use the text $\boldsymbol{x}$ to predict the features $\boldsymbol{z}$.** Specifically, we can train a logistic regression classifier to compute $\text{p}(z_k \mid \boldsymbol{x})$, for each $k \in \{1, 2, \ldots, K_z\}$.
2. **Use the features $\boldsymbol{z}$ to predict the label $y$.** Again, we can train a logistic classifier to compute $\text{p}(y \mid \boldsymbol{z})$. On test data, $\boldsymbol{z}$ is unknown, so we use the probabilities $\text{p}(\boldsymbol{z} \mid \boldsymbol{x})$ from the first layer as the features.

This setup is shown in Figure 2.1, which describes the proposed classifier in a **computation graph**: the text features $\boldsymbol{x}$ are connected to the middle layer $\boldsymbol{z}$, which in turn is connected to the label $y$.

[1] I will use "deep learning" and "neural networks" interchangeably.

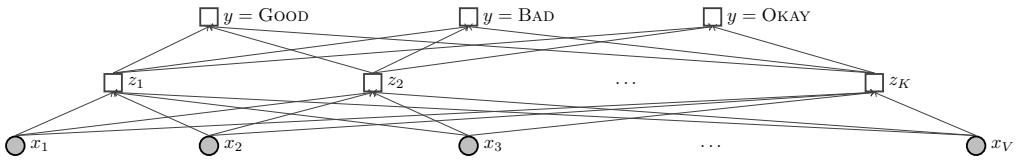



Figure 2.1: A feedforward neural network. Shaded circles indicate observed features, usually words; squares indicate nodes in the computation graph, which are computed from the information carried over the incoming arrows.

Since each $z_k \in \{0, 1\}$, we can treat $\text{p}(z_k \mid \boldsymbol{x})$ as a binary classification problem, using binary logistic regression:

$$\Pr(z_k = 1 \mid \boldsymbol{x}; \boldsymbol{\Theta}^{(x \to z)}) = \sigma(\boldsymbol{\theta}_k^{(x \to z)} \cdot \boldsymbol{x}) = (1 + \exp(-\boldsymbol{\theta}_k^{(x \to z)} \cdot \boldsymbol{x}))^{-1}, \quad [2.1]$$

where $\sigma(\cdot)$ is the **sigmoid** function (shown in Figure 2.2), and the matrix $\boldsymbol{\Theta}^{(x \to z)} \in \mathbb{R}^{K_z \times V}$ is constructed by stacking the weight vectors for each $z_k$,

$$\boldsymbol{\Theta}^{(x \to z)} = [\boldsymbol{\theta}_1^{(x \to z)}, \boldsymbol{\theta}_2^{(x \to z)}, \dots, \boldsymbol{\theta}_{K_z}^{(x \to z)}]^\top. \quad [2.2]$$

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

At the output layer, we use the multi-class logistic regression probability,

$$\Pr(y = j \mid \boldsymbol{z}; \boldsymbol{\Theta}^{(z \to y)}, \boldsymbol{b}) = \frac{\exp(\boldsymbol{\theta}_j^{(z \to y)} \cdot \boldsymbol{z} + b_j)}{\sum_{j' \in \mathcal{Y}} \exp(\boldsymbol{\theta}_{j'}^{(z \to y)} \cdot \boldsymbol{z} + b_{j'})}, \quad [2.3]$$

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

$$\boldsymbol{\Theta}^{(z \to y)} = [\boldsymbol{\theta}_1^{(z \to y)}, \boldsymbol{\theta}_2^{(z \to y)}, \dots, \boldsymbol{\theta}_{K_y}^{(z \to y)}]^\top. \quad [2.4]$$

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

$$\text{p}(\boldsymbol{y} \mid \boldsymbol{z}; \boldsymbol{\Theta}^{(z \to y)}, \boldsymbol{b}) = \text{SoftMax}(\boldsymbol{\Theta}^{(z \to y)} \boldsymbol{z} + \boldsymbol{b}), \quad [2.5]$$

where element $j$ in the output of the **SoftMax** function is computed as in Equation 2.3.

We have now defined a multilayer classifier, which can be summarized as,

$$\text{p}(\boldsymbol{z} \mid \boldsymbol{x}; \boldsymbol{\Theta}^{(x \to z)}) = \sigma(\boldsymbol{\Theta}^{(x \to z)} \boldsymbol{x}) \quad [2.6]$$

$$\text{p}(y \mid \boldsymbol{z}; \boldsymbol{\Theta}^{(z \to y)}, \boldsymbol{b}) = \text{SoftMax}(\boldsymbol{\Theta}^{(z \to y)} \boldsymbol{z} + \boldsymbol{b}), \quad [2.7]$$

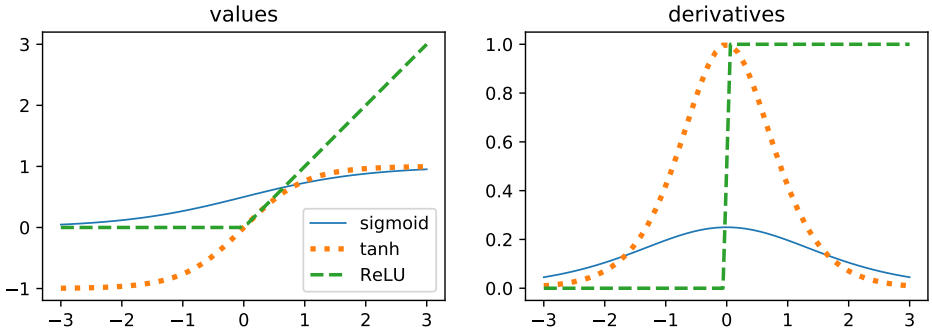



Figure 2.2: The sigmoid, tanh, and ReLU activation functions

where $\sigma(\cdot)$ is now applied **elementwise** to the vector of inner products,

$$\sigma(\mathbf{\Theta}^{(x\rightarrow z)}\boldsymbol{x}) = [\sigma(\boldsymbol{\theta}_1^{(x\rightarrow z)} \cdot \boldsymbol{x}), \sigma(\boldsymbol{\theta}_2^{(x\rightarrow z)} \cdot \boldsymbol{x}), \ldots, \sigma(\boldsymbol{\theta}_{K_z}^{(x\rightarrow z)} \cdot \boldsymbol{x})]^\top. \quad [2.8]$$

Now the hidden features $\boldsymbol{z}$ are never observed, even in the training data. We can still construct the architecture in Figure 2.1. Instead of predicting $y$ from a discrete vector of predicted values $\boldsymbol{z}$, we use the probabilities $\sigma(\boldsymbol{\theta}_k \cdot \boldsymbol{x})$. The resulting classifier is barely changed:

$$\boldsymbol{z} = \sigma(\mathbf{\Theta}^{(x\rightarrow z)}\boldsymbol{x}) \quad [2.9]$$

$$\text{p}(y \mid \boldsymbol{x}; \mathbf{\Theta}^{(z\rightarrow y)}, \boldsymbol{b}) = \text{SoftMax}(\mathbf{\Theta}^{(z\rightarrow y)}\boldsymbol{z} + \boldsymbol{b}). \quad [2.10]$$

We have now defined a classification model that predicts the label $y \in \mathcal{Y}$ from the base features $\boldsymbol{x}$, through a"hidden layer" $\boldsymbol{z}$. This is a **feedforward neural network**.[2]

## 2.2 Designing neural networks

This feedforward neural network can be generalized in a number of ways.

### 2.2.1 Activation functions

If the hidden layer is viewed as a set of latent features, then the sigmoid function represents the extent to which each of these features is "activated" by a given input. However, we can relax this view of the hidden layer, and regard it more generally as a nonlinear transformation of the input. This opens the door to many other activation functions, some of which are shown in Figure 2.2. At the moment, the choice of activation functions is more art than science, but a few points can be made about the most popular varieties:

[2]The architecture is sometimes called a **multilayer perceptron**, but this is misleading, because each layer is not a perceptron as defined in Algorithm 3.

- 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),

$$\text{ReLU}(a) = \begin{cases} a, & a \geq 0 \\ 0, & \text{otherwise.} \end{cases} \quad [2.11]$$

  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),

$$\text{Leaky-ReLU}(a) = \begin{cases} a, & a \geq 0 \\ .0001a, & \text{otherwise.} \end{cases} \quad [2.12]$$

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, with very small gradients on their own parameters. Several other activation functions are reviewed by Goodfellow et al. (2016), who recommend ReLU as the "default option."

### 2.2.2 Network structure

Deep networks stack up several hidden layers, $\boldsymbol{z}^{(1)}, \boldsymbol{z}^{(2)}, \ldots$, where each $\boldsymbol{z}^{(d)}$ is the input to the next layer, $\boldsymbol{z}^{(d+1)}$. As the total number of nodes in the network increases, so does its capacity to learn complex functions of the input. For a fixed number of nodes, an architectural decision is whether to emphasize width (large $K_z$ at each layer) or depth (many layers). At present, this tradeoff is not well understood.[3]

[3]There is a universal approximation theorem that states that with 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

It is also possible to "short circuit" a hidden layer, by propagating information directly from the input to the next higher level of the network. This is the idea behind **residual networks**, which propagate information directly from the input to the subsequent layer (He et al., 2016),

$$\boldsymbol{z} = f(\boldsymbol{\Theta}^{(x \rightarrow z)}\boldsymbol{x}) + \boldsymbol{x}, \tag{2.13}$$

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 $f(\boldsymbol{\Theta}^{(x \rightarrow z)}\boldsymbol{x})$ and $\boldsymbol{x}$:

$$\boldsymbol{t} = \sigma(\Theta^{(t)}\boldsymbol{x} + \boldsymbol{b}^{(t)}) \tag{2.14}$$

$$\boldsymbol{z} = \boldsymbol{t} \odot f(\boldsymbol{\Theta}^{(x \rightarrow z)}\boldsymbol{x}) + (\mathbf{1} - \boldsymbol{t}) \odot \boldsymbol{x}, \tag{2.15}$$

where $\odot$ refers to an elementwise vector product, and $\mathbf{1}$ is a column vector of ones. Gating is also used in the **long short-term memory (LSTM)**, which is discussed in chapter 5. Residual and highway connections address a problem with deep architectures: repeated application of a nonlinear activation function can make it difficult to learn the parameters of the lower levels of the network, which are too distant from the supervision signal.

### 2.2.3 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** loss function,

$$-\mathcal{L} = -\sum_{i=1}^{N} \log \mathrm{p}(y^{(i)} \mid \boldsymbol{x}^{(i)}; \Theta). \tag{2.16}$$

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

This loss can be written alternatively as follows:

$$\tilde{y}_j \triangleq \Pr(y = j \mid \boldsymbol{x}^{(i)}; \Theta) \tag{2.17}$$

$$-\mathcal{L} = -\sum_{i=1}^{N} \boldsymbol{e}_{y^{(i)}} \cdot \log \tilde{\boldsymbol{y}} \tag{2.18}$$

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

small non-zero error; see section 6.4.1 of Goodfellow et al. (2016) for a survey of these theoretical results. However, 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 say anything about whether it is possible to *learn* the function using gradient-based optimization.

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:

$$\Psi(y; \boldsymbol{x}^{(i)}, \Theta) = \boldsymbol{\theta}_y^{(z \to y)} \cdot \boldsymbol{z} + b_y \quad [2.19]$$

$$\ell_{\text{MARGIN}}(\Theta; \boldsymbol{x}^{(i)}, y^{(i)}) = \max_{y \neq y^{(i)}} \left(1 + \Psi(y; \boldsymbol{x}^{(i)}, \Theta) - \Psi(y^{(i)}; \boldsymbol{x}^{(i)}, \Theta)\right). \quad [2.20]$$

In regression problems, the output is a scalar or vector (see § 3.1.2). For these problems, a typical loss function is the squared error $(y - \hat{y})^2$ or squared norm $||\boldsymbol{y} - \hat{\boldsymbol{y}}||_2^2$.

### 2.2.4 Inputs and lookup layers

In text classification, the input layer $\boldsymbol{x}$ can refer to a bag-of-words vector, where $x_j$ is the count of word $j$. The input to the hidden unit $z_k$ is then $\sum_{j=1}^{V} \theta_{j,k}^{(x \to z)} x_j$, and word $j$ is represented by the vector $\boldsymbol{\theta}_j^{(x \to z)}$. This vector is sometimes described as the **embedding** of word $j$, and can be learned from unlabeled data, using techniques discussed in chapter 13. The columns of $\boldsymbol{\Theta}^{(x \to z)}$ are each $K_z$-dimensional word embeddings.

In chapter 1, we encountered another view of text documents, as a sequence of word tokens, $w_1, w_2, \ldots, w_M$. To implement this view in a neural network, represent each word token $w_m$ with a one-hot vector, $\boldsymbol{e}_{w_m} \in \mathbb{R}^V$. The matrix-vector product $\boldsymbol{\Theta}^{(x \to z)} \boldsymbol{e}_{w_m}$ returns the embedding of word $w_m$. The complete document can represented by horizontally concatenating these one-hot vectors, $\mathbf{W} = [\boldsymbol{e}_{w_1}, \boldsymbol{e}_{w_2}, \ldots, \boldsymbol{e}_{w_M}]$, and the bag-of-words representation can be recovered from the matrix-vector product $\mathbf{W}\mathbf{1}$, which simply sums over the tokens $m = \{1, 2, \ldots, M\}$. The matrix product $\boldsymbol{\Theta}^{(x \to z)} \mathbf{W}$ contains the horizontally concatenated embeddings of each word in the document, which will be useful as the starting point for **convolutional neural networks** (§ 2.4). This is sometimes called a **lookup layer**, because the first step is to lookup the embeddings for each word in the input text.

## 2.3 Learning neural networks

The feedforward network in Figure 2.1 can now be written in a more general form,

$$\boldsymbol{z} \leftarrow f(\boldsymbol{\Theta}^{(x \to z)} \boldsymbol{x}^{(i)}) \quad [2.21]$$

$$\tilde{\boldsymbol{y}} \leftarrow \text{SoftMax}\left(\boldsymbol{\Theta}^{(z \to y)} \boldsymbol{z} + \boldsymbol{b}\right) \quad [2.22]$$

$$\ell^{(i)} \leftarrow -\boldsymbol{e}_{y^{(i)}} \cdot \log \tilde{\boldsymbol{y}}, \quad [2.23]$$

where $f$ is an elementwise activation function, such as $\sigma$ or ReLU.

We now consider how to estimate the parameters: $\mathbf{\Theta}^{(x\rightarrow z)}$, $\mathbf{\Theta}^{(z\rightarrow y)}$ and $\boldsymbol{b}$. As in chapter 1, we will focus on online gradient-based optimization to incrementally minimize a local loss function with respect to each parameter. The simplest such algorithm is stochastic gradient descent,

$$\boldsymbol{b} \leftarrow \boldsymbol{b} - \eta^{(t)} \nabla_{\boldsymbol{b}} \ell^{(i)} \tag{2.24}$$

$$\boldsymbol{\theta}_k^{(z\rightarrow y)} \leftarrow \boldsymbol{\theta}_k^{(z\rightarrow y)} - \eta^{(t)} \nabla_{\boldsymbol{\theta}_k^{(z\rightarrow y)}} \ell^{(i)} \tag{2.25}$$

$$\boldsymbol{\theta}_k^{(x\rightarrow z)} \leftarrow \boldsymbol{\theta}_k^{(x\rightarrow z)} - \eta^{(t)} \nabla_{\boldsymbol{\theta}_k^{(x\rightarrow z)}} \ell^{(i)}, \tag{2.26}$$

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

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

$$\nabla_{\boldsymbol{\theta}_k^{(z\rightarrow y)}} \ell^{(i)} = \left[ \frac{\partial \ell^{(i)}}{\partial \theta_{k,1}^{(z\rightarrow y)}}, \frac{\partial \ell^{(i)}}{\partial \theta_{k,2}^{(z\rightarrow y)}}, \ldots, \frac{\partial \ell^{(i)}}{\partial \theta_{k,K_y}^{(z\rightarrow y)}} \right]^\top \tag{2.27}$$

$$\frac{\partial \ell^{(i)}}{\partial \theta_{k,j}^{(z\rightarrow y)}} = - \frac{\partial}{\partial \theta_{k,j}^{(z\rightarrow y)}} \left( \boldsymbol{\theta}_{y^{(i)}}^{(z\rightarrow y)} \cdot \boldsymbol{z} - \log \sum_{y \in \mathcal{Y}} \exp \boldsymbol{\theta}_y^{(z\rightarrow y)} \cdot \boldsymbol{z} \right) \tag{2.28}$$

$$= \left( \Pr(y = j \mid \boldsymbol{z}; \mathbf{\Theta}^{(z\rightarrow y)}, \boldsymbol{b}) - \delta\left(j = y^{(i)}\right) \right) z_k, \tag{2.29}$$

where $\delta\left(j = y^{(i)}\right)$ is a function that returns one when $j = y^{(i)}$, and zero otherwise. The gradient $\nabla_{\boldsymbol{b}} \ell^{(i)}$ is similar to Equation 2.29.

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

$$\frac{\partial \ell^{(i)}}{\partial \theta_{n,k}^{(x\rightarrow z)}} = \frac{\partial \ell^{(i)}}{\partial z_k} \frac{\partial z_k}{\partial \theta_{n,k}^{(x\rightarrow z)}} \tag{2.30}$$

$$= \frac{\partial \ell^{(i)}}{\partial z_k} \frac{\partial f(\boldsymbol{\theta}_k^{(x\rightarrow z)} \cdot \boldsymbol{x})}{\partial \theta_{n,k}^{(x\rightarrow z)}} \tag{2.31}$$

$$= \left( \frac{\partial \ell^{(i)}}{\partial z_k} \right) \times f'(\boldsymbol{\theta}_k^{(x\rightarrow z)} \cdot \boldsymbol{x}) \times x_n, \tag{2.32}$$

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

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

$$\frac{\partial \ell^{(i)}}{\partial \theta_{n,k}^{(x\rightarrow z)}} = \left(\frac{\partial \ell^{(i)}}{\partial z_k}\right) \times \sigma(\boldsymbol{\theta}_k^{(x\rightarrow z)} \cdot \boldsymbol{x}) \times (1 - \sigma(\boldsymbol{\theta}_k^{(x\rightarrow z)} \cdot \boldsymbol{x})) \times x_n \quad [2.33]$$

$$= \left(\frac{\partial \ell^{(i)}}{\partial z_k}\right) \times z_k \times (1 - z_k) \times x_n. \quad [2.34]$$

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

- If the negative log-likelihood $\ell^{(i)}$ does not depend much on $z_k$, $\frac{\partial \ell^{(i)}}{\partial z_k} \rightarrow 0$, then it doesn't matter how we compute $z_k$, and so $\frac{\partial \ell^{(i)}}{\partial \theta_{n,k}^{(x\rightarrow z)}} \rightarrow 0$.
- If $z_k$ is near 1 or 0, then the curve of the sigmoid function (Figure 2.2) is nearly flat, and changing the inputs will make little local difference. The term $z_k \times (1 - z_k)$ is maximized at $z_k = \frac{1}{2}$, where the slope of the sigmoid function is steepest.
- If $x_n = 0$, then it does not matter how we set the weights $\theta_{n,k}^{(x\rightarrow z)}$, so $\frac{\partial \ell^{(i)}}{\partial \theta_{n,k}^{(x\rightarrow z)}} = 0$.

### 2.3.1 Backpropagation

In the equations above, the value $\frac{\partial \ell^{(i)}}{\partial z_k}$ is reused in the derivatives with respect to each $\theta_{n,k}^{(x\rightarrow z)}$. It should therefore be computed once, and then cached. Furthermore, we should only compute any derivative once we have already computed all of the necessary "inputs" demanded by the chain rule of differentiation. This combination of sequencing, caching, and differentiation is known as **backpropagation**. It can be generalized to any directed acyclic **computation graph**.

A computation graph is a declarative representation of a computational process. At each node $t$, compute a value $v_t$ by applying a function $f_t$ to a (possibly empty) list of parent nodes, $\boldsymbol{\pi}_t$. For example, in a feedforward network with one hidden layer, there are nodes for the inputs $\boldsymbol{x}^{(i)}$, the hidden layer $\boldsymbol{z}$, the predicted output $\hat{\boldsymbol{y}}$, and the parameters $\{\boldsymbol{\Theta}^{(x\rightarrow z)}, \boldsymbol{\Theta}^{(z\rightarrow y)}, \boldsymbol{b}\}$. During training, there is also a node for the observed label $y^{(i)}$ and the loss $\ell^{(i)}$. Various types of nodes play different roles during learning:

- *Variables* include the *inputs* $\boldsymbol{x}$, the *hidden nodes* $\boldsymbol{z}$, the outputs $\boldsymbol{y}$, and the loss function. Inputs are variables that do not have parents. Backpropagation computes the gradients with respect to all variables except the inputs, but does not update the variables during learning.
- *Parameters* include the weights and offsets. They do not have parents, and they are updated during learning.

**Algorithm 6** General backpropagation algorithm. In the computation graph $G$, every node contains a function $f_t$ and a set of inputs $\boldsymbol{\pi}_t$; the inputs to the graph are $\boldsymbol{x}^{(i)}$. [todo: needs work]

**procedure** BACKPROP($G = \{f_t, \boldsymbol{\pi}_t\}_{t=1}^T\}, \boldsymbol{x}^{(i)}$)
  $v_{t(n)} \leftarrow x_n^{(i)}$ for all $n$ and associated computation nodes $t(n)$.
  **for** $t \in$ TOPOLOGICALSORT($G$) **do** ▷ Forward pass: compute value at each node
    **if** $|\pi_t| > 0$ **then**
      $v_t \leftarrow f_t(v_{\pi_{t,1}}, v_{\pi_{t,2}}, \ldots, v_{\pi_{t,N_t}})$
  $g_{\text{objective}} = 1$ ▷ Backward pass: compute gradients at each node
  **for** $t \in$ REVERSE(TOPOLOGICALSORT($G$)) **do**
    $g_t \leftarrow \sum_{t':t \in \boldsymbol{\pi}_{t'}} g_{t'} \frac{\partial f_{t'}(v_{\pi_{t',1}}, v_{\pi_{t',2}}, \ldots, v_{\pi_{t',N_{t'}}})}{\partial v_t}$
  **return** $\{g_1, g_2, \ldots, g_T\}$

- The *objective* is not the parent of any other node. Backpropagation begins by computing the gradient with respect to this node.

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 an input to node $t'$, then $t < t'$. This means that the values $\{v_t\}_{t=1}^T$ can be computed in a single forward pass. The reverse of the topological sort is used to compute gradients: each gradient $g_t$ is computed from the gradients of nodes to which $t$ is an input, implementing the chain rule of differentiation. The general backpropagation algorithm for computation graphs is shown in Algorithm 6, and illustrated in Figure 2.3.

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 end users need only specify the feed-forward computation, and the gradients necessary for learning can be obtained automatically. There are many software libraries that perform automatic differentiation on computation graphs, such as `Torch` (Collobert et al., 2011), `TensorFlow` (Abadi et al., 2016), and `DyNet` (Neubig et al., 2017). One important distinction between these libraries is whether they support **dynamic computation graphs**, in which the structure of the computation graph varies across instances. Static computation graphs are compiled in advance, and can be applied to fixed-dimensional data, such as bag-of-words vectors. In many natural language processing problems, it is desirable to represent the input as a sequence or even a tree structure, resulting in a variable computation graph at each instance.

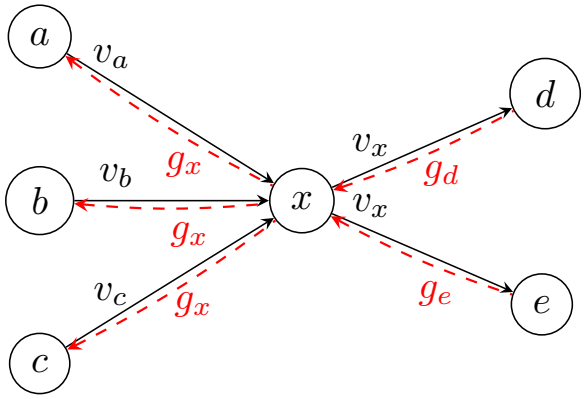



Figure 2.3: Backpropagation at a single node $x$ in the computation graph. The values of the predecessors $v_a, v_b, v_c$ are the inputs to $x$, which computes $v_x$, and passes it on to the successors $d$ and $e$. The gradients at the successors $g_d$ and $g_e$ are passed back to $x$, where they are incorporated into the gradient $g_x$, which is then passed back to the predecessors $a, b$, and $c$.

### 2.3.2 Regularization and dropout

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

$$L = \sum_{i=1}^{N} \ell^{(i)} + \lambda_{z\rightarrow y}||\boldsymbol{\Theta}^{(z\rightarrow y)}||_F^2 + \lambda_{x\rightarrow z}||\boldsymbol{\Theta}^{(x\rightarrow z)}||_F^2, \quad [2.35]$$

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

Another approach to controlling model complexity is **dropout**, which involves randomly setting some computation nodes to zero during training (Srivastava et al., 2014). For example, in the feedforward network, on each training instance, with probability $\rho$ we set each input $x_n$ and each hidden layer node $z_k$ to zero. Srivastava et al. (2014) recommend $\rho = 0.5$ for hidden units, and $\rho = 0.2$ for input units. Dropout is also incorporated in the gradient computation, so if node $z_k$ is dropped, then none of the weights $\boldsymbol{\theta}_k^{(x\rightarrow z)}$ will be updated for this instance. Dropout prevents the network from learning to depend too much on any one feature or hidden node, and prevents **feature co-adaptation**, in which a hidden unit is only useful in combination with one or more other hidden units. Dropout is

a special case of **feature noising**, which can also involve adding Gaussian noise to inputs or hidden units (Holmstrom and Koistinen, 1992). Wager et al. (2013) show that dropout is approximately equivalent to "adaptive" $L_2$ regularization, with a separate regularization penalty for each feature.

### 2.3.3 *Learning theory

Chapter 1 emphasized the importance of **convexity** for learning: for convex objectives, the global optimum can be found efficiently. The negative log-likelihood and hinge loss are convex functions of the parameters of the output layer. However, the output of a feedforward network is generally not a convex function of the parameters of the input layer, $\mathbf{\Theta}^{(x\rightarrow z)}$. Feedforward networks can be viewed as function composition, where each layer is a function that is applied to the output of the previous layer. Convexity is generally not preserved in the composition of two convex functions — and furthermore, "squashing" activation functions like tanh and sigmoid are not convex.

Here is a more intuitive explanation for why the objective of a neural network with one or more hidden layers is not a convex function of its inputs. Apply a permutation $\pi$ to the elements in the hidden layer, from $\boldsymbol{z} = [z_1, z_2, \ldots, z_{K_z}]$ to $\tilde{\boldsymbol{z}} = [z_{\pi(1)}, z_{\pi(2)}, \ldots, z_{\pi(K_z)}]$. This corresponds to applying $\pi$ to the rows of $\mathbf{\Theta}^{(x\rightarrow z)}$ and the columns of $\mathbf{\Theta}^{(z\rightarrow y)}$, resulting in permuted parameter matrices $\mathbf{\Theta}_\pi^{(x\rightarrow z)}$ and $\mathbf{\Theta}_\pi^{(z\rightarrow y)}$. As long as this permutation is applied consistently, the loss will be identical, $L(\mathbf{\Theta}) = L(\mathbf{\Theta}_\pi)$: it is *invariant* to this permutation. However, the loss of the linear combination $L(\alpha\mathbf{\Theta} + (1-\alpha)\mathbf{\Theta}_\pi)$ will generally not be identical to the loss under $\mathbf{\Theta}$ or its permutations. If $L(\mathbf{\Theta})$ is better than the loss at any points in the immediate vicinity, and if $L(\mathbf{\Theta}) = L(\mathbf{\Theta}_\pi)$, then the loss function does not satisfy the definition of convexity (see § 1.3). One of the exercises asks you to prove this more rigorously.

In practice, the existence of multiple optima is not necessary problematic, if all such optima are permutations of the sort described in the previous paragraph. In contrast, "bad" local optima are better than their neighbors, but much worse than the global optimum. Fortunately, in large feedforward neural networks, most local optima are nearly as good as the global optimum (Choromanska et al., 2015), which helps to explain why backpropagation works. More generally, a **critical point** is one at which the gradient is zero (this is also called a **stationary point**). Critical points may be local optima, but they may also be **saddle points**, which are local minima in some directions, but local *maxima* in other directions. For example, the equation $x_1^2 - x_2^2$ has a saddle point at $\boldsymbol{x} = (0, 0)$.[4] 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

[4]Thanks to Rong Ge's blogpost for this example, `http://www.offconvex.org/2016/03/22/saddlepoints/`

to zero. However, the noise introduced by stochastic gradient descent (and also 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).

### 2.3.4 Tricks

Getting neural networks to work effectively 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 are also problematic. 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),

$$\theta_{i,j} \sim U\left[-\frac{\sqrt{6}}{\sqrt{d_{\text{in}}(n)+d_{\text{out}}(n)}}, \frac{\sqrt{6}}{\sqrt{d_{\text{in}}(n)+d_{\text{out}}(n)}}\right], \tag{2.36}$$

[2.37]

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

$$\theta_{i,j} \sim N(0, \sqrt{2/d_{\text{in}}(n)}) \tag{2.38}$$

Rather than initializing the weights independently, it can be beneficial to initialize each layer jointly as an **orthonormal matrix**, ensuring that $\mathbf{\Theta}^\top\mathbf{\Theta} = \mathbf{I}$ (Saxe et al., 2014). Orthonormal matrices preserve the norm of the input, so that $||\mathbf{\Theta}\boldsymbol{x}|| = ||\boldsymbol{x}||$, 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:

$$a_{i,j} \sim N(0,1) \quad [2.39]$$

$$\mathbf{A} = \{a_{i,j}\}_{i=1,j=1}^{d_{\text{in}}(j),d_{\text{out}}(j)} \quad [2.40]$$

$$\mathbf{U}, \mathbf{S}, \mathbf{V}^\top = \text{SVD}(\mathbf{A}) \quad [2.41]$$

$$\mathbf{\Theta}^{(j)} \leftarrow \mathbf{U}. \quad [2.42]$$

The matrix $\mathbf{U}$ contains the **singular vectors** of $\mathbf{A}$, and is guaranteed to be orthonormal. For more on singular value decomposition, see chapter 13.

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

**Clipping and normalizing the gradients** As already discussed, the magnitude of the gradient can pose problems for learning: too large, and learning can diverge, with successive updates thrashing between increasingly extreme values; too small, and learning can grind to a halt. Several heuristics have been proposed to address this issue.

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

$$\text{CLIP}(\tilde{\boldsymbol{g}}) = \begin{cases} \boldsymbol{g} & ||\hat{\boldsymbol{g}}|| < \tau \\ \frac{\tau}{||\boldsymbol{g}||}\boldsymbol{g} & \text{otherwise.} \end{cases} \quad [2.43]$$

- 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 § 1.5.2). For example, in a feedforward network with one hidden layer, batch normalization would tranform the inputs to the hidden layer as follows:

$$\boldsymbol{\mu}^{(\mathcal{B})} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \boldsymbol{x}^{(i)} \quad [2.44]$$

$$\boldsymbol{s}^{(\mathcal{B})} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} (\boldsymbol{x}^{(i)} - \boldsymbol{\mu}^{(\mathcal{B})})^2 \quad [2.45]$$

$$\overline{\boldsymbol{x}}^{(i)} = (\boldsymbol{x}^{(i)} - \boldsymbol{\mu}^{(\mathcal{B})}) / \sqrt{\boldsymbol{s}^{(\mathcal{B})}}. \quad [2.46]$$

  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:

$$\boldsymbol{a} = \boldsymbol{\Theta}^{(x \to z)} \boldsymbol{x} \quad [2.47]$$

$$\mu = \frac{1}{K_z} \sum_{k=1}^{K_z} a_k \quad [2.48]$$

$$s = \frac{1}{K_z} \sum_{k=1}^{K_z} (a_k - \mu)^2 \quad [2.49]$$

$$\boldsymbol{z} = (\boldsymbol{a} - \mu)/\sqrt{s}. \quad [2.50]$$

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

**Online optimization** The trend towards deep learning has spawned a cottage industry of online optimization algorithms, which attempt to improve on stochastic gradient descent. **AdaGrad** was reviewed in § 1.5.2; 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,

$$v_j^{(t)} = \beta v_j^{(t-1)} + (1 - \beta) g_{t,j}^2, \quad [2.51]$$

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 § 1.2.2, can help to avoid overfitting, by terminating training after reaching a plateau in the performance on a heldout validation set.

## 2.4 Convolutional neural networks

A basic weakness of the bag-of-words model is its inability to account for the ways in 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*. Similarly, computer vision faces the challenge of identifying the semantics of images from pixel features that are profoundly uninformative in isolation. An earlier generation of computer vision research focused on designing *filters* to aggregate local pixel-level features into more meaningful representations, such as edges and corners (e.g., Canny, 1987). Similarly, earlier NLP research attempted to capture multiword linguistic phenomena by hand-designed lexical patterns (Hobbs et al., 1997). In both cases, the output of the filters and patterns

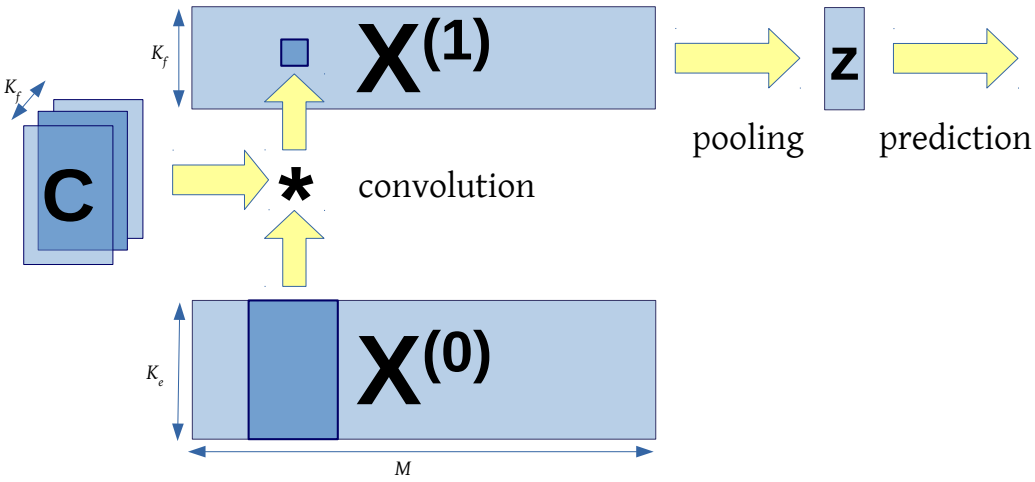



Figure 2.4: A convolutional neural network for text classification

could then act as base features in a linear classifier. But rather than designing these feature extractors by hand, a better approach is to learn them, using the magic of backpropagation. This is the idea behind **convolutional neural networks**.

Following § 2.2.4, define the base layer of a neural network as $\mathbf{X}^{(0)} = \mathbf{\Theta}^{(x \to z)}[\boldsymbol{e}_{w_1}, \boldsymbol{e}_{w_2}, \ldots, \boldsymbol{e}_{w_M}]$, where $\boldsymbol{e}_{w_m}$ 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 $\mathbf{C}^{(k)} \in \mathbb{R}^{K_e \times h}$. Convolution is indicated by the symbol $*$, and is defined,

$$\mathbf{X}^{(1)} = f(\boldsymbol{b} + \mathbf{C} * \mathbf{X}^{(0)}) \quad \Longrightarrow \quad x^{(1)}_{k,m} = f\left(b_k + \sum_{k'=1}^{K_e} \sum_{n=1}^{h} c^{(k)}_{k',n} \times x^{(0)}_{k',m+n-1}\right), \qquad [2.52]$$

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

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

$$\mathbf{C}^{(k)} = \begin{bmatrix} 0.5 & 1 & 0.5 \\ 0.5 & 1 & 0.5 \\ \ldots & \ldots & \ldots \\ 0.5 & 1 & 0.5 \end{bmatrix}, \qquad [2.53]$$

thereby representing trigram units like *not so unpleasant*. In **one-dimensional convolution**, each filter matrix $\mathbf{C}^{(k)}$ is constrained to have non-zero values only at row $k$ (Kalchbrenner et al., 2014).

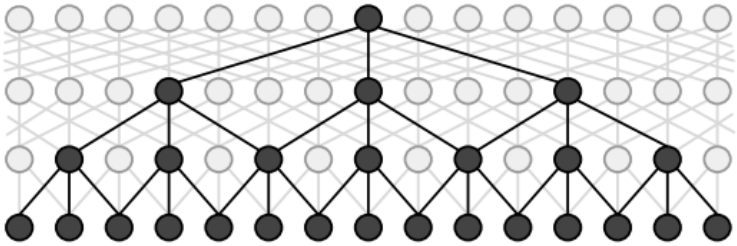


Figure 2.5: A dilated convolutional neural network captures progressively larger context through recursive application of the convolutional operator (Strubell et al., 2017) [todo: permission]

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

After $D$ convolutional layers, we obtain a matrix representation of the document $\mathbf{X}^{(D)} \in \mathbb{R}^{K_z \times M}$. 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, such as max-pooling (Collobert et al., 2011) or average-pooling,

$$\boldsymbol{z} = \text{MaxPool}(\mathbf{X}^{(D)}) \quad \Longrightarrow \quad z_k = \max\left(x_{k,1}^{(D)}, x_{k,2}^{(D)}, \ldots x_{k,M}^{(D)}\right) \tag{2.54}$$

$$\boldsymbol{z} = \text{AvgPool}(\mathbf{X}^{(D)}) \quad \Longrightarrow \quad z_k = \frac{1}{M}\sum_{m=1}^{M} x_{k,m}^{(D)}. \tag{2.55}$$

The vector $\boldsymbol{z}$ can now act as a layer in a feedforward network, culminating in a prediction $\hat{y}$ and a loss $\ell^{(i)}$. The setup is shown in Figure 2.4.

Just as in feedforward networks, the parameters $(\mathbf{C}^{(k)}, \boldsymbol{b}, \Theta)$ 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 $m$:

$$\frac{\partial z_k}{\partial x_{k,m}^{(D)}} = \begin{cases} 1, & x_{k,m}^{(D)} = \max\left(x_{k,1}^{(D)}, x_{k,2}^{(D)}, \ldots x_{k,M}^{(D)}\right) \\ 0, & \text{otherwise.} \end{cases} \tag{2.56}$$

The computer vision literature has produced a huge variety of convolutional architectures, and many of these bells and whistles can be applied to text data. One avenue for

improvement is more complex pooling operations, such as $k$-max pooling (Kalchbrenner et al., 2014), which returns a matrix of the $k$ largest values for each filter. Another innovation is the use of **dilated convolution** to build multiscale representations (Yu and Koltun, 2016). At each layer, the convolutional operator applied in *strides*, skipping ahead by $s$ steps after each feature. As we move up the hierarchy, each layer is $s$ times smaller than the layer below it, effectively summarizing the input. This idea is shown in Figure 2.5. Multi-layer convolutional networks can also be augmented with "shortcut" connections, as in the ResNet model from § 2.2.2 (Johnson and Zhang, 2017).

## Further reading

The deep learning textbook by Goodfellow et al. (2016) covers many of the topics in this chapter in more detail. For a comprehensive review of neural networks in natural language processing, see (Goldberg, 2017b). A seminal work on deep learning in natural language processing is the aggressively titled "Natural Language Processing (Almost) from Scratch", which uses convolutional neural networks to perform a range of language processing tasks (Collobert et al., 2011). This chapter focuses on feedforward and convolutional neural networks, but recurrent neural networks are one of the most important deep learning architectures for natural language processing. They are covered extensively in chapters 5 and 6.

The role of deep learning in natural language processing research has caused angst in some parts of the natural language processing research community (e.g., Goldberg, 2017a), especially as some of the more zealous deep learning advocates have argued that end-to-end learning from "raw" text can eliminate the need for linguistic constructs such as sentences, phrases, and even words (Zhang et al., 2015, originally titled *Text understanding from scratch*). These developments were surveyed by Manning (2016).

## Exercises

1. Prove that the softmax and sigmoid functions are equivalent when the number of possible labels is two. Specifically, for any $\mathbf{\Theta}^{(z \rightarrow y)}$ (omitting the offset $\boldsymbol{b}$ for simplicity), show how to construct a vector of weights $\boldsymbol{\theta}$ such that,

$$\text{SoftMax}(\mathbf{\Theta}^{(z \rightarrow y)} \boldsymbol{z})[0] = \sigma(\boldsymbol{\theta} \cdot \boldsymbol{z}). \quad [2.57]$$

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

$$f(x_1, x_2) = \begin{cases} -1, & x_1 = 1, x_2 = 1 \\ 1, & x_1 = 1, x_2 = 0 \\ 1, & x_1 = 0, x_2 = 1 \\ -1, & x_1 = 0, x_2 = 0 \end{cases}. \quad [2.58]$$

Your network should have a single output node which uses the Sign activation function. Use a single hidden layer, with ReLU activation functions. Describe all weights and offsets.

3. 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. Prove that gradient-based optimization cannot learn the desired function from this initialization.

4. 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 it is not possible to learn the desired function.

5. Consider a network in which: the base features are all binary, $\boldsymbol{x} \in \{0, 1\}^M$; the hidden layer activation function is sigmoid, $z_k = \sigma(\boldsymbol{\theta}_k \cdot \boldsymbol{x})$; and the initial weights are sampled independently from a standard normal distribution, $\theta_{j,k} \sim N(0, 1)$.

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

$$\Pr(\sigma(\boldsymbol{\theta}_k \cdot \boldsymbol{x}) \times (1 - \sigma(\boldsymbol{\theta}_k \cdot \boldsymbol{x})) < \alpha) \quad \geq \quad 2 \Pr(\sigma(\boldsymbol{\theta}_k \cdot \boldsymbol{x}) < \alpha), \quad [2.59]$$

     and relate this probability to the variance $V[\boldsymbol{\theta}_k \cdot \boldsymbol{x}]$.
   - Design an alternative initialization that removes this dependence.

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

$$\Big( ||\tilde{\Theta}^{(x \to z)} - \Theta^{(x \to z)}||_F^2 + ||\tilde{\Theta}^{(z \to y)} - \Theta^{(z \to y)}||_F^2 + ||\tilde{\boldsymbol{b}} - \boldsymbol{b}||_2^2 < \epsilon \Big)$$
$$\Rightarrow \Big( L(\tilde{\Theta}) > L(\Theta) \Big) \quad [2.60]$$

   Define the function $\pi$ as a permutation on the hidden units, as described in § 2.3.3, so that for any $\Theta$, $L(\Theta) = L(\Theta_\pi)$. Prove that if a feedforward network has a local optimum in the sense of Equation 2.60, then its loss is not a convex function of the parameters $\Theta$, using the definition of convexity from § 1.3

# Chapter 3

# Linguistic applications of classification

Having learned some techniques for classification, we will now see how they can be applied to some classical problems in natural language technology. Later in this chapter, we discuss some of the design decisions involved in text classification, as well as evaluation practices.

## 3.1 Sentiment and opinion analysis

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

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

- Tweets containing happy emoticons can be marked as positive, sad emoticons as

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

negative (Read, 2005; Pak and Paroubek, 2010).

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

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

The problem becomes more tricky for short documents, such as single-sentence reviews or social media posts. In these documents, linguistic issues like **negation** and **irrealis** (Polanyi and Zaenen, 2006) — events that are hypothetical or otherwise non-factual — can make bag-of-words classification ineffective. Consider the following examples:

(3.1) That's not bad for the first day.

(3.2) This is not the worst thing that can happen.

(3.3) It would be nice if you acted like you understood.

(3.4) There is no reason at all to believe that the polluters are suddenly going to become reasonable. (Wilson et al., 2005)

(3.5) This film should be brilliant. The actors are first grade. Stallone plays a happy, wonderful man. His sweet wife is beautiful and adores him. He has a fascinating gift for living life fully. It sounds like a great plot, **however**, the film is a failure. (Pang et al., 2002)

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

$$\langle \textit{that's}, \textit{not} \rangle, \langle \textit{not}, \textit{bad} \rangle, \langle \textit{bad}, \textit{for} \rangle, \ldots \quad [3.1]$$

Bigrams can handle relatively straightforward cases, such as when an adjective is immediately negated; trigrams would be required to extend to larger contexts (e.g., *not the worst*). But it should be clear that this approach will not scale to the more complex examples, such as (3.4) and (3.5). More sophisticated solutions try to account for the syntactic structure of the sentence (Wilson et al., 2005; Socher et al., 2013), or apply more complex classifiers such as **convolutional neural networks** (Kim, 2014), which are described in chapter 2.

### 3.1.1 Related problems

**Subjectivity** Closely related to sentiment analysis is **subjectivity detection**, which requires identifying the parts of a text that express subjective opinions, as well as other non-factual content such speculation and hypotheticals (Riloff and Wiebe, 2003). This can be done by treating each sentence as a separate document, and then applying a bag-of-words classifier: indeed, Pang and Lee (2004) do exactly this, using a training set consisting of (mostly) subjective sentences gathered from movie reviews, and (mostly) objective sentences gathered from plot descriptions. They augment this bag-of-words model with a graph-based algorithm that encourages nearby sentences to have the same subjectivity label.

**Stance classification** In debates, each participant takes a side: for example, advocating for or against adopting a vegetarian lifestyle or mandating free college education. The problem of stance classification involves identifying an author's position from the text of the argument. In some cases, there is training data available for each position, so that standard document classification techniques can be employed. In other cases, it suffices to classify each document as whether it is in support or opposition of the argument advanced by a previous document (Anand et al., 2011). In the most challenging case, there is no labeled data for any of the stances, so the only possibility is group documents that advocate the same position (Somasundaran and Wiebe, 2009). This is a form of **unsupervised learning**, and will be discussed in chapter 4.

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

(3.6) The vodka was good, but the meat was rotten.

(3.7) Go to Heaven for the climate, Hell for the company. (Mark Twain)

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

**Aspect-based opinion mining** seeks to identify the sentiment of the author of a review towards predefined aspects such as PRICE and SERVICE, or, in the case of (3.7), CLIMATE and COMPANY (Hu and Liu, 2004). If the aspects are not defined in advance, it may again be necessary to employ **unsupervised learning** methods to identify them (e.g., Branavan et al., 2009).

**Emotion classification** While sentiment analysis is framed in terms of positive and negative categories, psychologists generally regard **emotion** as more multifaceted. For example, Ekman (1992) argues that there are six basic emotions — happiness, surprise, fear, sadness, anger, and contempt — and that they are universal across human cultures. Alm et al. (2005) build a linear classifier for recognizing the emotions expressed in children's stories. The ultimate goal of this work was to improve text-to-speech synthesis, so that stories could be read with intonation that reflected the emotional content. They used bag-of-words features, as well as features capturing the story type (e.g., jokes, folktales), and structural features that reflect the position of each sentence in the story. The task is difficult: even human annotators frequently disagreed with each other, and the best classifiers achieved accuracy between 60-70%.

### 3.1.2 Alternative approaches to sentiment analysis

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

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

$$\hat{y} = \underset{r:\, \boldsymbol{\theta} \cdot \boldsymbol{x} \geq b_r}{\operatorname{argmin}}\, r, \qquad [3.2]$$

where $\boldsymbol{b} = [b_1 = -\infty, b_2, b_3, \ldots, b_K]$ is a vector of boundaries. Crammer and Singer (2001) show that it is possible to learn both the weights and boundaries simultaneously, using a perceptron-like algorithm.

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

$$\theta_{y,j} = \begin{cases} 1, & j \in \mathcal{L}_y \\ 0, & \text{otherwise}, \end{cases} \qquad [3.3]$$

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

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

## 3.2 Word sense disambiguation

Consider the the following headlines:

(3.8) *Iraqi head seeks arms*

(3.9) *Prostitutes appeal to Pope*

(3.10) *Drunk gets nine years in violin case*[2]

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

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

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

then choosing the correct sense from the inventory associated with the corresponding lemma.[3]

### 3.2.1 How many word senses?

Words (lemmas) may have many more than two senses. For example, the word *serve* would seem to have at least the following senses:

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

These sense distinctions are annotated in **WordNet** (`http://wordnet.princeton.edu`), a lexical semantic database for English. WordNet consists of roughly 100,000 **synsets**, which are groups of lemmas (or phrases) that are synonymous. An example synset is $\{chump^1, fool^2, sucker^1, mark^9\}$, where the superscripts index the sense of each lemma that is included in the synset: for example, there are at least eight other senses of *mark* that have different meanings, and are not part of this synset. A lemma is **polysemous** if it participates in multiple synsets.

WordNet plays an key role in setting the parameters of the word sense disambiguation problem, and in formalizing lexical semantic knowledge of English. (WordNets have been created for a few dozen other languages, at varying levels of detail.) Some have argued that WordNet's sense granularity is too fine (Ide and Wilks, 2006); more fundamentally, the premise that word senses can be differentiated in a task-neutral way has been criticized as linguistically naïve (Kilgarriff, 1997). One way of testing this question is to ask whether people tend to agree on the appropriate sense for example sentences: according to Mihalcea et al. (2004), people agree on roughly 70% of examples using WordNet senses; far better than chance, but perhaps less than we might like.

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

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

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

[4] Several of the examples are adapted from `WordNet` (Fellbaum, 2010)

- **hyponymy**: $x$ is a special case of $y$, e.g. RED-COLOR; the inverse relationship is **hypernymy**;
- **meronymy**: $x$ is a part of $y$, e.g., WHEEL-BICYCLE; the inverse relationship is **holonymy**.

Classification of these relations relations can be performed by searching for characteristic patterns between pairs of words, e.g., *X, such as Y*, which signals hyponymy (Hearst, 1992), or *X but Y*, which signals antonymy (Hatzivassiloglou and McKeown, 1997). Another approach is to analyze each term's **distributional statistics** (the frequency of its neighboring words). Such approaches are described in detail in chapter 13.

### 3.2.2 Word sense disambiguation as classification

How can we tell living *plants* from manufacturing *plants*? The key information often lies in the context:

(3.11) Town officials are hoping to attract new manufacturing plants through weakened environmental regulations.

(3.12) 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 treating each context as a pseudo-document. We can then construct a feature function for each potential sense $y$,

$$f(\langle \textit{plant}, \textit{The endangered plants play an} \ldots \rangle, y) = \\ \{\langle \textit{the}, y\rangle : 1, \langle \textit{endangered}, y\rangle : 1, \langle \textit{play}, y\rangle : 1, \langle \textit{an}, y\rangle : 1, \ldots\}$$

As in document classification, many of these features are irrelevant, but a few are very strong indicators. In this example, the context word *endangered* is a strong signal that the intended sense is biology rather than manufacturing. We would therefore expect a learning algorithm to assign high weight to $\langle \textit{endangered}, \text{BIOLOGY}\rangle$, and low weight to $\langle \textit{endangered}, \text{MANUFACTURING}\rangle$.[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.,

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

[5]The context bag-of-words can be also used be 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.

These **collocation features** give more information about the specific role played by each context word. This idea can be taken further by incorporating additional syntactic information about the grammatical role played by each context feature, such as the **dependency path** (see chapter 10).

After deciding on the features, we can train a classifier to predict the sense of each word. 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, annotated as part of the WordNet project (Fellbaum, 2010). SemCor annotations look like this:

(3.13) As of Sunday$^{1}_{n}$ night$^{1}_{n}$ there was$^{4}_{v}$ no word$^{2}_{n}$ ...,

with the superscripts indicating the annotated sense of each polysemous word.

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 **semisupervised** methods are particularly important for WSD (e.g., Yarowsky, 1995). These methods will be discussed in chapter 4. 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).

## 3.3 Design decisions for text classification

Text classification involves a number of design decisions. Some of these decisions, such as smoothing or regularization, are specific to the classifier. These decisions are described in the previous chapter. But even the construction of the feature vector itself involves a number of design decisions, and these decisions can be the most consequential for the classifier's performance.

### 3.3.1 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, 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 3.1: The output of four `nltk` tokenizers, applied to the string *Isn't Ahab, Ahab? ;)*

### Tokenization

The first subtask for constructing a bag-of-words vector is **tokenization**: converting the text from a sequence of characters to a sequence of **word tokens**. A simple approach is to define a subset of characters as whitespace, and then split the text on these tokens. However, whitespace-based tokenization is not ideal: we may want to split conjunctions like *isn't* and hyphenated phrases like *prize-winning* and *half-asleep*, and we likely want to separate words from commas and periods that immediately follow them. This suggests that a tokenizer should split on all non-alphanumeric characters, but we would prefer not to split abbreviations like *U.S.* and *Ph.D.* In languages with Roman scripts, tokenization is typically performed using regular expressions, with modules designed to handle each of these cases. For example, the `nltk` package includes a number of tokenizers; the outputs of four of the better-known tokenizers are shown in Figure 3.1. Social media researchers have found that emoticons and other forms of orthographic variation pose new challenges for tokenization, leading to the development of special purpose tokenizers to handle these phenomena (O'Connor et al., 2010).

Tokenization is a language-specific problem, and each language poses unique challenges. For example, Chinese does not include spaces between words, nor any other consistent orthographic markers of word boundaries. A "greedy" approach is to scan the input for character substrings that are in a predefined lexicon. However, Xue et al. (2003) notes that this can be ambiguous, since many character sequences could be segmented in multiple ways. Instead, he trains a classifier to determine whether each Chinese character, or **hanzi**, is a word boundary. More advanced sequence labeling methods for word segmentation are discussed in § 7.4. Similar problems can occur in languages with alphabetic scripts, such as German, which does not include whitespace in compound nouns, yielding examples such as *Freundschaftsbezeigungen* and *Dilettantenaufdringlichkeiten* [todo: ask German speaker for better examples]. As Twain (1997) argues, "*These things are not words, they are alphabetic processions.*" Social media raises similar problems for English and other languages, with hashtags such as *#TrueLoveInFourWords* requiring 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 |

Figure 3.2: The outputs of the Porter (1980) and Lancaster (Paice, 1990) stemmers

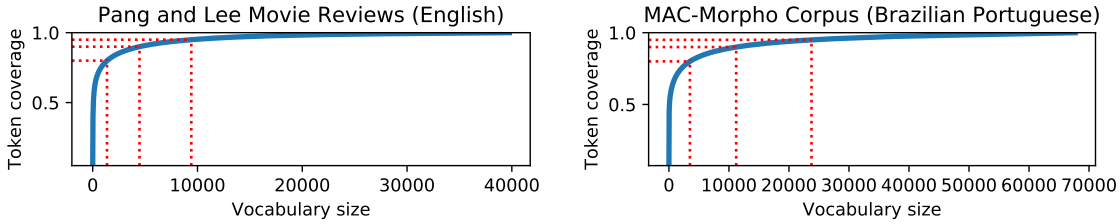
### Normalization

After splitting the text into tokens, the next question is which tokens are really distinct. Is it necessary to distinguish *great*, *Great*, and GREAT? Sentence-initial capitalization may be irrelevant to the classification task. The elimination of case distinctions will result in a smaller vocabulary, and thus smaller feature vectors. However, case distinctions might be relevant in some situations: for example, *apple* is a delicious pie filling, while *Apple* is a company dedicated to marketing proprietary dongles and power adapters. For Latin script, case conversion can be performed using unicode string libraries. Many scripts do not have case distinctions (e.g., the Devanagari script used for South Asian languages, the Thai alphabet, and Japanese kana), and case conversion for all scripts may not be available in every programming environment.

Case conversion is an example of **normalization**, which refers to string transformations that remove distinctions that are felt to be irrelevant (Sproat et al., 2001). Other normalizations include the standardization of numbers (e.g., *1,000* to *1000*) and dates (e.g., *August 11, 2015* to *2015/11/08*). Depending on the application, it may even be worthwhile to convert all numbers and dates to special tokens, `!NUM` and `!DATE`. Social media features orthographic phenomena such as expressive lengthening (e.g., *cooooool*), which may also be normalized (Aw et al., 2006; Yang and Eisenstein, 2013). Similarly, historical texts feature spelling variations which may be normalized to a standard form (Baron and Rayson, 2008).

A more extreme form of normalization is to eliminate **inflectional affixes**, such as the *-ed* and *-s* suffixes in English. On this view, *bike*, *bikes*, *biking*, and *biked* all refer to the same underlying concept, so they should be grouped into a single feature. A **stemmer** is a program for eliminating affixes, usually by applying a series of regular expression substitutions. Character-based stemming algorithms are necessarily approximate, as shown in Figure 3.2: the Lancaster stemmer incorrectly identifies *-ers* as an inflectional suffix of *sisters* (by analogy to *fix/fixers*), and both stemmers incorrectly identify *-s* as a suffix of *this* and *Williams*. Fortunately, even inaccurate stemming can improve bag-of-words classification models, by merging related strings and thereby reducing the vocabulary size.

Accurately handling irregular orthography requires word-specific rules. **Lemmatizers** are systems that identify the underlying lemma of a given wordform. They must avoid the over-generalization errors of the stemmers in Figure 3.2, and also handle more complex




(a) Movie review data in English

(b) News articles in Brazilian Portuguese

Figure 3.3: 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.

transformations, such as *geese→goose*. The output of the WordNet lemmatizer is shown in the final line of Figure 3.2. Both stemming and lemmatization are language-specific: an English stemmer or lemmatizer is of little use on a text written in another language. The discipline of **morphology** relates to the study of word-internal structure, and is described in more detail in § 8.1.2.

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

### 3.3.2 How many words?

Limiting the size of the feature vector reduces the memory footprint of the resulting models, and increases the speed of prediction. Normalization can help to play this role, but a more direct approach is simply to limit the vocabulary to the $N$ most frequent words in the dataset. For example, in the `movie-reviews` dataset provided with `nltk` (originally from Pang et al., 2002), there are 39,768 word types, and 1.58M tokens. As shown in Figure 3.3a, 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 3.3b). 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 well-regularized discriminative classifiers. As with normalization, stopword filtering is more important for unsupervised problems, such as term-based document retrieval; in this case, matching a *to* or *the* in the search query offers little information that the resulting document will meet the goals of the user's search.

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).

### 3.3.3 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: two *failure*s may be worse 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: $f_j(\boldsymbol{x}, y) \in \{0, 1\}$. 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.

## 3.4 Evaluating classifiers

In any supervised machine learning application, it is critical to reserve a held-out test set, and use this data for only one purpose: to evaluate the overall accuracy of a single classifier. Using this data more than once would cause the estimated accuracy to be overly optimistic, because the classifier would be customized to this data, and would not perform as well as on unseen data in the future. It is usually necessary to set hyperparameters or

perform feature selection, so you may need to construct a **tuning** or **development set** for this purpose, as discussed in § 1.1.5.

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

$$\text{acc}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{N} \sum_{i}^{N} \delta(y^{(i)} = \hat{y}). \quad [3.4]$$

If you've ever taken an exam, it was probably graded by accuracy. Why are other metrics necessary? The main reason is **class imbalance**. Suppose we were building a classifier to detect whether a electronic health record (EHR) described symptoms of a rare disease, which appear in only 1% of all documents in the dataset. A classifier that reports $\hat{y} = -1$ for all documents would achieve 99% accuracy, but would be practically useless. We need metrics that are capable of detecting the classifier's ability to discriminate between classes, even when the distribution is skewed.

One solution would be to build a **balanced test set**, in which 50% of documents are positive. But this would mean throwing away 98% of the original dataset! Furthermore, the detection threshold itself might be a design consideration: in health-related applications, we might prefer a very sensitive classifier, which returned a positive prediction if there is even a small chance that $y^{(i)} = +1$. In other applications, a positive result might trigger a costly action, so we would prefer a classifier that only makes positive predictions when absolutely certain. We need additional metrics to capture these characteristics.

### 3.4.1 Precision, recall, and $F$-measure

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

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

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

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

Classifiers that make a lot of false positive errors are too sensitive; classifiers that make a lot of false negative errors are not sensitive enough. We can capture these two behaviors

through two additional metrics, **recall** and **precision**:

$$\text{recall}(\boldsymbol{y}, \hat{\boldsymbol{y}}, k) = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad [3.5]$$

$$\text{precision}(\boldsymbol{y}, \hat{\boldsymbol{y}}, k) = \frac{\text{TP}}{\text{TP} + \text{FP}}. \quad [3.6]$$

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

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

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

$$F\text{-measure}(\boldsymbol{y}, \hat{\boldsymbol{y}}, k) = \frac{2rp}{r + p}, \quad [3.7]$$

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), it is necessary to combine the $F$-measure across each class. **Macro** $F$**-measure** is the average $F$-measure across several classes,

$$\text{Macro-}F(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} F\text{-measure}(\boldsymbol{y}, \hat{\boldsymbol{y}}, k) \quad [3.8]$$

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

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

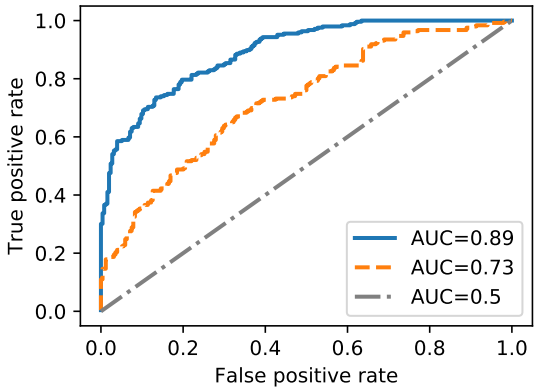



Figure 3.4: ROC curves for three classifiers of varying discriminative power, measured by AUC (area under the curve)

### 3.4.2 Threshold-free metrics

In binary classification problems, it is possible to trade off between recall and precision by adding a constant "threshold" to the output of the scoring function. This makes it possible to trace out a curve, where each point indicates the performance at a single threshold. In the **receiver operating characteristic (ROC)** curve,[7], the x-axis indicates the **false positive rate**, $\frac{\text{FP}}{\text{FP+TN}}$, and the y-axis indicates the recall, or **true positive rate**. A perfect classifier attains perfect recall without any false positives, tracing a "curve" from the origin (0,0) to the upper left corner (0,1), and then to (1,1). In expectation, a non-discriminative classifier traces a diagonal line from the origin (0,0) to the upper right corner (1,1). Real classifiers tend to fall between these two extremes. Examples are shown in Figure 3.4.

The ROC curve can be summarized in a single number by taking its integral, the **area under the curve (AUC)**. The AUC has an intuitive interpretation as the probability that a randomly-selected positive example will be assigned a higher score by the classifier than a randomly-selected negative example. Thus, a perfect classifier has AUC $= 1$ (all positive examples score higher than all negative examples); a random 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.

[7]The 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.

### 3.4.3 Classifier comparison and statistical significance

Building NLP systems 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 feature sets discussed in § 3.2.2. **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 an effect of equal or greater magnitude when the null hypothesis (that the classifiers are equally accurate) is true. For example, we write $p < .05$ when 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 the test set, and that $c_1$ is correct on $k$ of those instances. Under the null hypothesis that the classifiers are equally accurate, we would expect $k/N$ to be roughly equal to $1/2$. As $N$ increases, $k/N$ should be increasingly close to $1/2$. These properties are captured by the **binomial distribution**, which is a probability over counts of binary random variables.

[8] Other 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.

[9] Statistical 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.

[10] A 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. 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 explain from first principles.

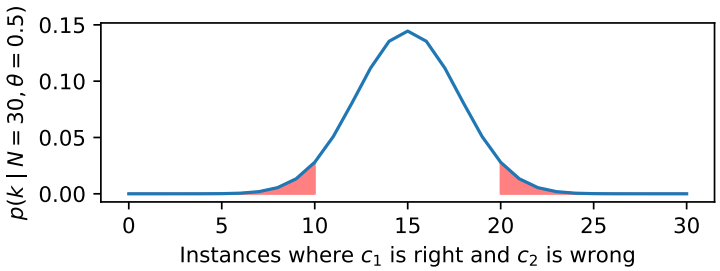



Figure 3.5: Probability mass function for the binomial distribution. The pink highlighted areas represent the cumulative probability for a significance test on an observation of $k = 10$ and $N = 30$.

We write $k \sim \text{Binom}(\theta, N)$ to indicate that $k$ is drawn from a binomial distribution, with parameter $N$ indicating the number of random "draws", and $\theta$ indicating the probability of "success" on each draw. The **probability mass function** (PMF) of the binomial distribution is,

$$\text{p}_{\text{Binom}}(k; N, \theta) = \binom{N}{k} \theta^k (1-\theta)^{N-k}, \quad [3.9]$$

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

Under the null hypothesis, $\theta = \frac{1}{2}$: when the classifiers disagree, they are each equally likely to be right. Now suppose that among $N$ disagreements, $c_1$ is correct only $k < \frac{N}{2}$ times. The probability of $c_1$ being correct $k$ or fewer times is the **one-tailed p-value**, 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 3.5. This **cumulative probability** is computed as a sum over all values $i \leq k$,

$$\Pr_{\text{Binom}}\left(\text{count}(\hat{y}_2^{(i)} = y^{(i)} \neq \hat{y}_1^{(i)}) \leq k; N, \theta = \frac{1}{2}\right) = \sum_{i=0}^{k} \text{p}_{\text{Binom}}\left(i; N, \theta = \frac{1}{2}\right). \quad [3.10]$$

The one-tailed p-value applies only to the asymmetric null hypothesis that $c_1$ is at least as accurate as $c_2$. To test the **two-tailed** null hypothesis that $c_1$ and $c_2$ are equally accurate, we would take the sum of one-tailed $p$-values, where the second term is computed from the right tail of Figure 3.5. The binomial distribution is symmetric, so this can be computed by simply doubling the one-tailed p-value.

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

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

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

### *Randomized testing

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

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

confidence interval from the mean and variance.

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

### 3.4.4 *Multiple comparisons

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

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 $\alpha$. 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 $\alpha$: 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.

## 3.5 Building datasets

For many classification tasks of interest, no labeled data exists. If you want to build a classifier, you must first build a dataset of your own. This includes selecting a set of documents or instances to annotate, and then performing the annotations.

In many cases, the scope of the dataset is determined by the application: if you want to build a system to classify electronic health records, then you must work with a corpus of records of the type that your classifier will encounter when deployed. In other cases, the goal is to build a system that will work across a broad range of documents. In this case, it is best to have a **balanced** corpus, with contributions from many styles and genres. For example, the Brown corpus draws from texts ranging from government documents to romance novels (Francis, 1964), and the Google Web Treebank includes annotations for five "domains" of web documents: question answers, emails, newsgroups, reviews, and blogs (Petrov and McDonald, 2012).

### 3.5.1 Metadata as labels

Annotation is difficult and time-consuming, and most people would rather avoid it. Luckily, it is sometimes possible to exploit existing metadata to obtain the desired labels. For example, reviews are often accompanied by a numerical rating, which can be converted into a classification label (see § 3.1). Similarly, the nationalities of social media users can be estimated from their profiles (Dredze et al., 2013) or even the time zones of their posts (Gouws et al., 2011). More ambitiously, we may try to classify the political affiliations of social media profiles based on their social network connections to politicians and major political parties (Rao et al., 2010).

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

### 3.5.2 Labeling data

In many cases, there is no way to get ground truth labels other than manual annotation. Good annotations should satisfy several criteria: they should be **expressive** enough to capture the phenomenon of interest; they should be **replicable**, meaning that another annotator or team of annotators would produce very similar annotations if given the same data; and they should be **scalable**, so that they can be produced relatively quickly. Hovy and Lavid (2010) propose a structured procedure for obtaining annotations that meet these criteria, which is summarized below.

1. **Determine what the annotations are to include**. This is usually based on some theory of the underlying phenomenon: for example, if the goal is to produce annotations about the emotional state of a document's author, one should start with an theoretical account of the types or dimensions of emotion (e.g., Mohammad and Turney, 2013). At this stage, the tradeoff between expressiveness and scalability should be considered: a full instantiation of the underlying theory might be too costly to annotate at scale, so reasonable approximations should be considered.

2. Optionally, one may **design or select a software tool to support the annotation**

**effort**. Existing general-purpose annotation tools include BRAT (Stenetorp et al., 2012) and MMAX2 (Müller and Strube, 2006).

3. **Formalize the instructions for the annotation task**. To the extent that the instructions are not explicit, the resulting annotations will depend on the intuitions of the annotators. These intuitions may not be shared by other annotators, or by the users of the annotated data. Therefore explicit instructions are critical to ensuring the annotations are replicable and usable by other researchers.

4. **Perform a pilot annotation** of a small subset of data, with multiple annotators for each instance. This will give a preliminary assessment of both the replicability and scalability of the current annotation instructions. Metrics for computing the rate of agreement are described below. Manual analysis of specific disagreements should help to clarify the instructions, and may lead to modifications of the annotation task itself. For example, if two labels are commonly conflated by annotators, it may be best just to merge them.

5. After finalizing the annotation protocol and instructions, the main annotation effort can begin. Some if not all of the instances should receive multiple annotations, so that inter-annotator agreement can be computed. In some annotation projects, instances receive many annotations, which are then aggregated into a "consensus" label (e.g., Danescu-Niculescu-Mizil et al., 2013). However, if the annotations are time-consuming or require significant expertise, it may be preferable to maximize scalability by obtaining multiple annotations for only a small subset of examples.

6. Compute and report inter-annotator agreement, and release the data. In some cases, the raw text data cannot be released, usually 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.

#### Measuring inter-annotator agreement

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

**Cohen's Kappa** is widely used for quantifying the agreement on discrete labeling tasks (Cohen, 1960; Carletta, 1996),[11]

$$\kappa = \frac{\text{agreement} - E[\text{agreement}]}{1 - E[\text{agreement}]}. \quad [3.11]$$

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

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

$$E[\text{agreement}] = \sum_k \hat{\Pr}(Y = k)^2, \quad [3.12]$$

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

### Crowdsourcing

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

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

## Exercises

1. As noted in § 3.3.3, words tend to appear in clumps, with subsequent occurrences of a word being more probable. More concretely, if word $j$ has probability $\phi_{y,j}$ of appearing in a document with label $y$, then the probability of two appearances $(x_j^{(i)} = 2)$ is greater than $\phi_{y,j}^2$.

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

   $$\Pr(w = j \mid y = 1) > \Pr(w = j \mid y = 0). \quad [3.13]$$

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

2. Prove that F-measure is never greater than the arithmetic mean of recall and precision, $\frac{r+p}{2}$. Your solution should also show that F-measure is equal to $\frac{r+p}{2}$ iff $r = p$.

3. Given a binary classification problem in which the probability of the "positive" label is equal to $\alpha$, what is the expected $F$-measure of a random classifier which ignores the data, and selects $\hat{y} = +1$ with probability $\frac{1}{2}$? (Assume that $\text{p}(\hat{y}) \perp \text{p}(y)$.) What is the expected $F$-measure of a classifier that selects $\hat{y} = +1$ with probability $\alpha$ (also independent of $y^{(i)}$)? Depending on $\alpha$, which random classifier will score better?

4. Suppose that binary classifiers $c_1$ and $c_2$ disagree on $N = 30$ cases, and that $c_1$ is correct in $k = 10$ of those cases.

   - Write a program that uses primitive functions such as $\exp$ and factorial to compute the **two-tailed** $p$-value — you may use an implementation of the "choose" function if one is avaiable. Verify your code against the output of a library for computing the binomial test or the binomial CDF, such as `scipy.stats.binom` in Python.
   - Then use a randomized test to try to obtain the same $p$-value. In each sample, draw from a binomial distribution with $N = 30$ and $\theta = \frac{1}{2}$. Count the fraction of samples in which $k \leq 10$. This is the one-tailed $p$-value; double this to compute the two-tailed $p$-value.
   - Try this with varying numbers of bootstrap samples: $M \in \{100, 1000, 5000, 10000\}$. For $M = 100$ and $M = 1000$, run the test 10 times, and plot the resulting $p$-values.
   - Finally, perform the same tests for $N = 70$ and $k = 25$.

5. SemCor 3.0 is a labeled dataset for word sense disambiguation. You can download it,[12] or access it in `nltk.corpora.semcor`.

   Choose a word that appears at least ten times in SemCor (*find*), and annotate its WordNet senses across ten randomly-selected examples, without looking at the ground truth. Use online WordNet to understand the definition of each of the senses.[13] Have a partner do the same annotations, and compute the raw rate of agreement, expected chance rate of agreement, and Cohen's kappa.

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 randomly-selected 400 reviews as a test set.

   Download a sentiment lexicon, such as the one currently available from Bing Liu, `https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html`. Tokenize the data, and classify each document as positive iff it has more positive sentiment words than negative sentiment words. Compute the accuracy and $F$-measure on detecting positive reviews on the test set, using this lexicon-based classifier.

   Then train a discriminative classifier (averaged perceptron or logistic regression) on the training set, and compute its accuracy and $F$-measure on the test set.

   Determine whether the differences are statistically significant, using two-tailed hypothesis tests: Binomial for the difference in accuracy, and bootstrap for the difference in macro-$F$-measure.

The remaining problems will require you to build a classifier and test its properties. Pick a multi-class text classification dataset, such as RCV1[14]). Divide your data into training (60%), development (20%), and test sets (20%), if no such division already exists.

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

8. Compare the following tokenization algorithms:

   - Whitespace, using a regular expression

[12] e.g., `https://github.com/google-research-datasets/word_sense_disambigation_corpora` or `http://globalwordnet.org/wordnet-annotated-corpora/`

[13] http://wordnetweb.princeton.edu/perl/webwn

[14] `http://www.ai.mit.edu/projects/jmlr/papers/volume5/lewis04a/lyrl2004_rcv1v2_README.htm`

- Penn Treebank
- Split input into five-character units, regardless of whitespace or punctuation

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

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

10. Identify the best combination of vocabulary filtering, tokenization, and stemming from the previous three problems. Apply this preprocessing to the test set, and compute the test set accuracy and Macro-$F$-measure. Compare against a baseline system that applies no vocabulary filtering, whitespace tokenization, and no stemming.

    Use the binomial test to determine whether your best-performing system is significantly more accurate than the baseline.

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

# Chapter 4

# Learning without supervision

So far we've assumed the following setup:

- A **training set** where you get observations $\boldsymbol{x}^{(i)}$ and labels $y^{(i)}$
- A **test set** where you only get observations $\boldsymbol{x}^{(i)}$

If you never get any labeled instances, is it possible to learn anything? This scenario is known as **unsupervised learning**, and we will see that indeed it is possible to learn about the underlying structure of unlabeled observations. We will also explore some related scenarios: **semi-supervised learning**, in which only some instances are labeled, and **domain adaptation**, in which the training data differs from the data on which the trained system will be deployed.

## 4.1 Unsupervised learning

To motivate unsupervised learning, consider the problem of word sense disambiguation (§ 3.2). Our goal is to classify each instance of a word, such as *bank* into a sense,

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

It is difficult to obtain sufficient training data for word sense disambiguation, because even a large corpus will contain only a few instances of all but the most common words. Unsupervised learning — **word sense induction** — would be highly desirable for this problem.

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

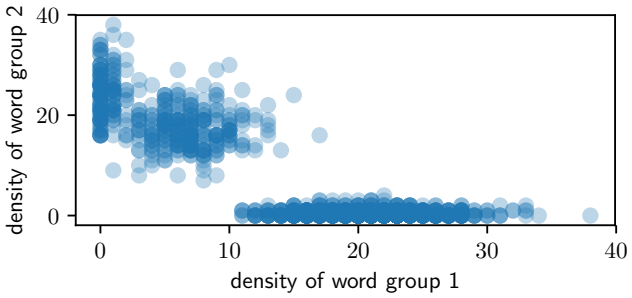



Figure 4.1: Counts of words from two different context groups

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

1. *deposits, interest, regulation*

2. *sediment, river, sand*

Suppose we were to scatterplot each instance of *bank* on a graph, so that the $x$-axis is the density of words in group 1, and the $y$-axis is the density of words in group 2. In such a plot, shown in Figure 4.1, two or more "blobs" might emerge, and these blobs correspond to the different senses of *bank*. These blobs represent the underlying structure of the unlabeled data.

While this points towards a solution, we don't know the groupings of context words in advance. We have to apply the same idea without predefined word groupings, by working in a high-dimensional space, with one dimension for every context word. Although we can't plot this space on paper, the idea is the same: try to identify the underlying structure of $\boldsymbol{x}$, such that there are a few clusters of points, each of which is internally coherent.

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

These examples show that even without labels, we can often find structure in data. In two dimensions, groupings of data points can be identified by eye, but this is not possible in the high dimensional scenarios that we face in natural language processing. Instead, we will employ **clustering** algorithms to find this structure automatically.

**Algorithm 8** $K$-means clustering algorithm

**procedure** $K$-MEANS($\boldsymbol{x}_{1:N}, K$)
  **for** $i \in 1 \ldots N$ **do** ▷ initialize cluster memberships
    $z^{(i)} \leftarrow \textsf{RandomInt}(1, K)$
  **repeat**
    **for** $k \in 1 \ldots K$ **do** ▷ recompute cluster centers
      $\boldsymbol{\nu}_k \leftarrow \frac{1}{\delta(z^{(i)}=k)} \sum_{i=1}^{N} \delta(z^{(i)} = k)\boldsymbol{x}^{(i)}$
    **for** $i \in 1 \ldots N$ **do** ▷ reassign instances to nearest clusters
      $z^{(i)} \leftarrow \text{argmin}_k \, ||\boldsymbol{x}^{(i)} - \boldsymbol{\nu}_k||^2$
  **until** converged
  **return** $\{z^{(i)}\}$ ▷ return cluster assignments

### 4.1.1 $K$-means clustering

Clustering algorithms assign each data point to a discrete cluster, $z_i \in 1, 2, \ldots K$. A classical clustering algorithm is **$K$-means**, an iterative algorithm that maintains a cluster assignment for each instance and a central location for each cluster. $K$-means iterates between updates to the assignments and the centers:

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

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

This is formalized in Algorithm 8. The term $||\boldsymbol{x}^{(i)} - \boldsymbol{\nu}||^2$ refers to the squared Euclidean norm, $\sum_n (x_n^{(i)} - \nu_n)^2$.

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

$$\boldsymbol{\nu}_k = \frac{1}{\sum_{i=1}^{N} q^{(i)}(k)} \sum_{i=1}^{N} q^{(i)}(k)\boldsymbol{x}^{(i)}. \quad [4.1]$$

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

### 4.1.2 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,

$$\log \mathrm{p}(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\phi}, \boldsymbol{\mu}) = \sum_{i=1}^{N} \log \Big( \mathrm{p}(\boldsymbol{x}^{(i)} \mid y^{(i)}; \boldsymbol{\phi}) \times \mathrm{p}(y^{(i)}; \boldsymbol{\mu}) \Big) \tag{4.2}$$

Suppose we 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 $\boldsymbol{z}$, we compute a distribution over the observed instances $\boldsymbol{x}$:

$$\log \mathrm{p}(\boldsymbol{x}; \boldsymbol{\phi}, \boldsymbol{\mu}) = \sum_{i=1}^{N} \log \mathrm{p}(\boldsymbol{x}^{(i)}; \boldsymbol{\phi}, \boldsymbol{\mu}) \tag{4.3}$$

$$= \sum_{i=1}^{N} \log \sum_{z=1}^{K} \mathrm{p}(\boldsymbol{x}^{(i)}, z; \boldsymbol{\phi}, \boldsymbol{\mu}) \tag{4.4}$$

$$= \sum_{i=1}^{N} \log \sum_{z=1}^{K} \mathrm{p}(\boldsymbol{x}^{(i)} \mid z; \boldsymbol{\phi}) \times \mathrm{p}(z; \boldsymbol{\mu}). \tag{4.5}$$

To estimate the parameters $\boldsymbol{\phi}$ and $\boldsymbol{\mu}$, we can maximize the marginal likelihood in Equation 4.5. Why is this the right thing to maximize? If we don't have labels, discriminative learning is impossible — there's nothing to discriminate. So maximum likelihood is all we have.

When the labels are observed, we can estimate the parameters of the Naïve Bayes probability model separately for each label. But marginalizing over the labels couples these parameters, making direct optimization of $\log \mathrm{p}(\boldsymbol{x})$ intractable. We will approximate the log-likelihood by introducing an **auxiliary variable** $\boldsymbol{q}^{(i)}$, which is a distribution over the label set $\mathcal{Z} = \{1, 2, \ldots, K\}$. Our optimization procedure will alternate between updates to $\boldsymbol{q}$ and updates to the parameters $(\boldsymbol{\phi}, \boldsymbol{\mu})$. Thus, $\boldsymbol{q}^{(i)}$ plays the same role in EM as in soft $K$-means.

To derive the updates for this optimization, multiply the right side of Equation 4.5 by

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

$$\log \mathrm{p}(\boldsymbol{x}; \boldsymbol{\phi}, \boldsymbol{\mu}) = \sum_{i=1}^{M} \log \sum_{z=1}^{K} \mathrm{p}(\boldsymbol{x}^{(i)} \mid z; \boldsymbol{\phi}) \times \mathrm{p}(z; \boldsymbol{\mu}) \times \frac{q^{(i)}(z)}{q^{(i)}(z)} \tag{4.6}$$

$$= \sum_{i=1}^{M} \log \sum_{z=1}^{K} q^{(i)}(z) \times \mathrm{p}(\boldsymbol{x}^{(i)} \mid z; \boldsymbol{\phi}) \times \mathrm{p}(z; \boldsymbol{\mu}) \times \frac{1}{q^{(i)}(z)} \tag{4.7}$$

$$= \sum_{i=1}^{M} \log E_{\boldsymbol{q}^{(i)}} \left[ \frac{\mathrm{p}(\boldsymbol{x}^{(i)} \mid z; \boldsymbol{\phi}) \mathrm{p}(z; \boldsymbol{\mu})}{q^{(i)}(z)} \right], \tag{4.8}$$

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

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

$$\log \mathrm{p}(\boldsymbol{x}; \boldsymbol{\phi}, \boldsymbol{\mu}) \geq \sum_{i=1}^{N} E_{\boldsymbol{q}^{(i)}} \left[ \log \frac{\mathrm{p}(\boldsymbol{x}^{(i)} \mid z; \boldsymbol{\phi}) \mathrm{p}(z; \boldsymbol{\mu})}{q^{(i)}(z)} \right] \tag{4.9}$$

$$\mathcal{J} \triangleq \sum_{i=1}^{N} E_{\boldsymbol{q}^{(i)}} \left[ \log \mathrm{p}(\boldsymbol{x}^{(i)} \mid z; \boldsymbol{\phi}) + \log \mathrm{p}(z; \boldsymbol{\mu}) - \log q^{(i)}(z) \right] \tag{4.10}$$

$$= \sum_{i=1}^{N} E_{\boldsymbol{q}^{(i)}} \left[ \log \mathrm{p}(\boldsymbol{x}^{(i)}, z; \boldsymbol{\phi}, \boldsymbol{\mu}) \right] + H(\boldsymbol{q}^{(i)}) \tag{4.11}$$

We will focus on Equation 4.10, which is the lower bound on the marginal log-likelihood of the observed data, $\log \mathrm{p}(\boldsymbol{x})$. Equation 4.11 shows the connection to the information theoretic concept of **entropy**, $H(\boldsymbol{q}^{(i)}) = -\sum_z q^{(i)}(z) \log q^{(i)}(z)$, which measures the average amount of information produced by a draw from the distribution $q^{(i)}$. Returning to Equation 4.10, this bound is a function of two groups of arguments:

- the distributions $\boldsymbol{q}^{(i)}$ for each instance;
- the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\phi}$.

We will optimize with respect to each of these in turn, while holding the other fixed.

### The E-step

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

write out the expectation in the lower bound as a sum,

$$\mathcal{J} = \sum_{i=1}^{N} \sum_{z=1}^{K} q^{(i)}(z) \left[ \log \mathrm{p}(\boldsymbol{x}^{(i)} \mid z; \boldsymbol{\phi}) + \log \mathrm{p}(z; \boldsymbol{\mu}) - \log q^{(i)}(z) \right]. \tag{4.12}$$

As in Naïve Bayes, we have a "sum-to-one" constraint: in this case, $\sum_{z=1}^{K} q^{(i)}(z) = 1$. Once again, we incorporate this constraint into a Lagrangian:

$$\mathcal{J}_q = \sum_{i=1}^{N} \sum_{z=1}^{K} q^{(i)}(z) \left( \log \mathrm{p}(\boldsymbol{x}^{(i)} \mid z; \boldsymbol{\phi}) + \log \mathrm{p}(z; \mu) - \log q^{(i)}(z) \right) + \lambda^{(i)}(1 - \sum_{z=1}^{K} q^{(i)}(z)) \tag{4.13}$$

We then optimize by taking the derivative and setting it equal to zero:

$$\frac{\partial \mathcal{J}_q}{\partial q^{(i)}(z)} = \log \mathrm{p}(\boldsymbol{x}^{(i)} \mid z; \boldsymbol{\phi}) + \log \mathrm{p}(z; \boldsymbol{\theta}) - \log q^{(i)}(z) - 1 - \lambda^{(i)} \tag{4.14}$$

$$\log q^{(i)}(z) = \log \mathrm{p}(\boldsymbol{x}^{(i)} \mid z; \boldsymbol{\phi}) + \log \mathrm{p}(z; \mu) - 1 - \lambda^{(i)} \tag{4.15}$$

$$q^{(i)}(z) \propto \mathrm{p}(\boldsymbol{x}^{(i)} \mid z; \boldsymbol{\phi}) \times \mathrm{p}(z; \mu). \tag{4.16}$$

Since $\boldsymbol{q}^{(i)}$ is constrained to be a probability distribution, the normalized probability can be computed easily,

$$q^{(i)}(z) = \frac{\mathrm{p}(\boldsymbol{x}^{(i)} \mid z; \boldsymbol{\phi}) \times \mathrm{p}(z; \boldsymbol{\mu})}{\sum_{z'=1}^{K} \mathrm{p}(\boldsymbol{x}^{(i)} \mid z'; \boldsymbol{\phi}) \times \mathrm{p}(z'; \boldsymbol{\mu})} \tag{4.17}$$

$$= \mathrm{p}(z \mid \boldsymbol{x}^{(i)}; \boldsymbol{\phi}, \boldsymbol{\mu}). \tag{4.18}$$

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

### The M-step

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

$$\mathcal{J}_\phi = \sum_{i=1}^{N} \sum_{z=1}^{K} q^{(i)}(z) \left( \log \mathrm{p}(\boldsymbol{x}^{(i)} \mid z; \boldsymbol{\phi}) + \log \mathrm{p}(z; \mu) - \log q^{(i)}(z) \right) + \sum_{z=1}^{K} \lambda_z (1 - \sum_{j=1}^{V} \phi_{z,j}). \tag{4.19}$$

The conditional log-likelihood for the multinomial is,

$$\log \mathrm{p}(\boldsymbol{x}^{(i)} \mid z, \boldsymbol{\phi}) = C + \sum_{j=1}^{V} x_j \log \phi_{z,j}, \tag{4.20}$$

where $C$ is a constant with respect to $\boldsymbol{\phi}$. Setting the derivative of $\mathcal{J}_\phi$ equal to zero,

$$\frac{\partial \mathcal{J}_\phi}{\partial \phi_{z,j}} = \sum_{i=1}^{N} q^{(i)}(z) \times \frac{x_j^{(i)}}{\phi_{z,j}} - \lambda_z \tag{4.21}$$

$$\phi_{z,j} \propto \sum_{i=1}^{N} q^{(i)}(z) \times x_j^{(i)}. \tag{4.22}$$

Because $\boldsymbol{\phi}_z$ is constrained to be a probability distribution, the normalized probability can be computed easily,

$$\phi_{z,j} = \frac{\sum_{i=1}^{N} q^{(i)}(z) \times x_j^{(i)}}{\sum_{j'=1}^{V} \sum_{i=1}^{N} q^{(i)}(z) \times x_{j'}^{(i)}} = \frac{E_q\left[\text{count}(z,j)\right]}{\sum_{j'=1}^{V} E_q\left[\text{count}(z,j')\right]}, \tag{4.23}$$

where the counter $j \in \{1, 2, \ldots, V\}$ indexes over base features, such as words.

This update sets $\boldsymbol{\phi}_z$ equal to the relative frequency estimate of the **expected counts** under the distribution $\boldsymbol{q}$. As in supervised Naïve Bayes, we can smooth these counts by adding a constant $\alpha$. The update for $\boldsymbol{\mu}$ is similar: $\mu_z \propto \sum_{i=1}^{N} q^{(i)}(z) = E_q\left[\text{count}(z)\right]$, the expected frequency of cluster $z$. This probability can also be smoothed. In sum, the M-step is just like Naïve Bayes, but with expected counts rather than observed counts.

The multinomial likelihood $\mathrm{p}(\boldsymbol{x} \mid z)$ can be replaced with other probability distributions: for example, for continuous observations, a Gaussian distribution can be used. In some cases, there is no closed-form update to the parameters of the likelihood. One approach is to run gradient-based optimization at each M-step; another is to simply take a single step along the gradient step and then return to the E-step (Berg-Kirkpatrick et al., 2010).

### 4.1.3 EM as an optimization algorithm

Algorithms that alternate between updating various subsets of the parameters are called **coordinate ascent** algorithms. The objective function $\mathcal{J}$ is **biconvex**: it is separately convex in $\boldsymbol{q}$ and $\langle \boldsymbol{\mu}, \boldsymbol{\phi} \rangle$, but it is not jointly convex in all terms. In the coordinate ascent algorithm that we have defined, each step is guaranteed not to decrease $\mathcal{J}$. Thanks to this guarantee, we know that EM will converge towards a solution at which no nearby points yield further improvements. This solution is a **local optimum** — it is as good or

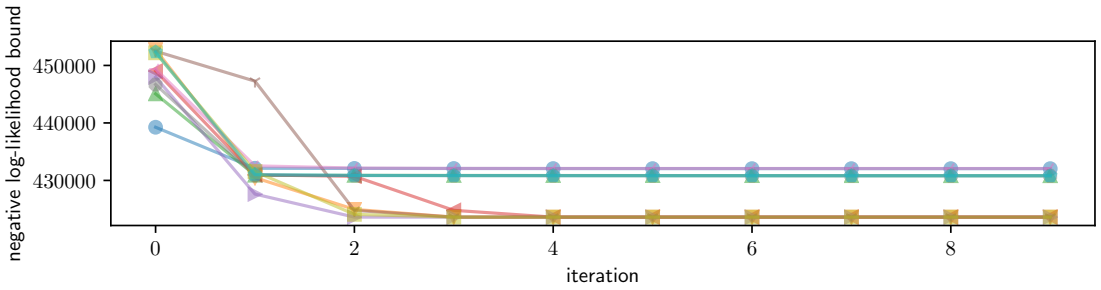



Figure 4.2: Sensitivity of expectation maximization to initialization. Each line shows the progress of optimization from a different random initialization.

better than any of its immediate neighbors, but is **not** guaranteed to be optimal among all possible configurations of $\langle \boldsymbol{q}, \boldsymbol{\mu}, \boldsymbol{\phi} \rangle$.

The fact that there is no guarantee of global optimality means that initialization is important: where you start can determine where you finish. To illustrate this point, Figure 4.2 shows the objective function for EM with ten different random initializations: while the objective function improves monotonically in each run, it converges to several different values.[1] For linear classifiers like logistic regression, we don't need to worry about initialization, because it won't affect the ultimate solution: we are guaranteed to reach the global minimum. (Initialization does matter for more complex supervised learning algorithms like multilayer neural networks.) Recent work on **spectral learning** has sought to obtain similar guarantees for unsupervised learning; this is briefly described in § 4.5.

In **hard EM**, each $\boldsymbol{q}^{(i)}$ distribution assigns probability of 1 to a single label $\hat{z}^{(i)}$, and zero probability to all others (Neal and Hinton, 1998). This is similar in spirit to $K$-means clustering, and can outperform standard EM in some cases (Spitkovsky et al., 2010). Another variant of coordinate ascent combines EM with stochastic gradient descent (SGD). In this case, we can do a local E-step at each instance $\boldsymbol{x}^{(i)}$, and then immediately make a gradient update to the parameters $\langle \boldsymbol{\mu}, \boldsymbol{\phi} \rangle$. This algorithm has been called **incremental expectation maximization** (Neal and Hinton, 1998) and **online expectation maximization** (Sato and Ishii, 2000; Cappé and Moulines, 2009), and is especially useful when there is no closed-form optimum for the likelihood $\text{p}(\boldsymbol{x} \mid z)$, and in online settings where new data is constantly streamed in (see Liang and Klein, 2009, for a comparison for online EM variants).

[1] The figure shows the bound on the **negative** log-likelihood, because optimization is typically framed as minimization rather than maximization.

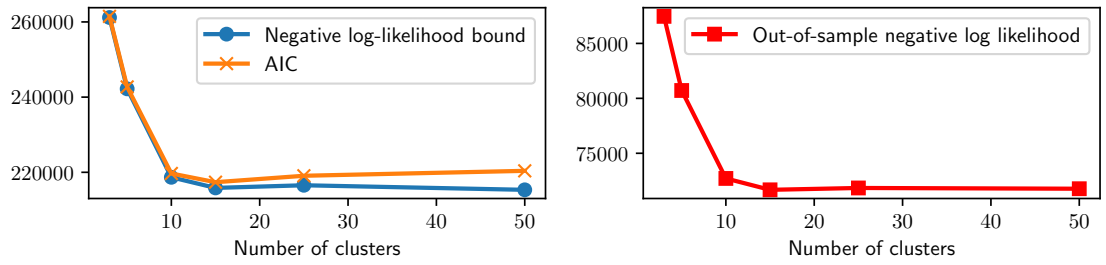



Figure 4.3: The negative log-likelihood and AIC for several runs of expectation maximization, on synthetic data. Although the data was generated from a model with $K = 10$, the optimal number of clusters is $\hat{K} = 15$, according to AIC and the heldout log-likelihood. The training set log-likelihood continues to improve as $K$ increases.

### 4.1.4 How many clusters?

So far, we have assumed that the number of clusters $K$ is given. In some cases, this assumption is valid. For example, a lexical semantic resource like WordNet might tell us the number of senses for a word. The number of clusters might also be a parameter for the user to tune: some readers want a coarse-grained clustering of news stories into three or four clusters, while others want a fine-grained clustering into twenty or more. But in many cases, we must choose $K$ automatically, with little extrinsic guidance.

One solution is to choose the number of clusters to maximize metric of clustering quality. The other parameters $\boldsymbol{\mu}$ and $\phi$ are chosen to maximize the log-likelihood bound $\mathcal{J}$, so this might seem a potential candidate for tuning $K$. However, $\mathcal{J}$ will never decrease with $K$: if it is possible to obtain a bound of $\mathcal{J}_K$ with $K$ clusters, then it is always possible to do at least as well with $K + 1$ clusters, by simply ignoring the additional cluster and setting its probability to zero in $\boldsymbol{q}$ and $\boldsymbol{\mu}$. It is therefore necessary to introduce a penalty for model complexity, so that fewer clusters are preferred. For example, the Akaike Information Crition (AIC; Akaike, 1974) is the linear combination of the number of parameters and the log-likelihood,

$$\text{AIC} = 2M - 2\mathcal{J}, \tag{4.24}$$

where $M$ is the number of parameters. In an expectation maximization clustering algorithm, $M = K \times V + K$. Since the number of parameters increases with the number of clusters $K$, the AIC may prefer more parsimonious models, even if they do not fit the data quite as well. Another choice is to maximize the **predictive likelihood** on heldout data $\boldsymbol{x}^{(h)}_{1:N_h}$. This data is not used to estimate the model parameters $\phi$ and $\boldsymbol{\mu}$, and so it is not the case that the likelihood on this data is guaranteed to increase with $K$. Figure 4.3 shows the negative log-likelihood on training and heldout data, as well as the AIC.

***Bayesian nonparametrics** An alternative approach is to treat the number of clusters as another latent variable. This requires statistical inference over a set of models with a variable number of clusters. This is not possible within the framework of expectation maximization, but there are several alternative inference procedures which can be applied, including **Markov Chain Monte Carlo (MCMC)**, which is briefly discussed in § 4.5 (for more details, see Chapter 25 of Murphy, 2012). Bayesian nonparametrics have been applied to the problem of unsupervised word sense induction, thus learning not only the word senses but also the number of senses per word (Reisinger and Mooney, 2010).

## 4.2 Applications of expectation-maximization

EM is not really an "algorithm" like, say, quicksort. Rather, it is a framework for learning with missing data. The recipe for using EM on a problem of interest is:

- Introduce latent variables $\boldsymbol{z}$, such that it is easy to write the probability $P(\mathcal{D}, \boldsymbol{z})$, where $\mathcal{D}$ is the observed data. It should also be easy to estimate the associated parameters, given knowledge of $\boldsymbol{z}$.
- Derive the E-step updates for $q(\boldsymbol{z})$, which is typically factored as $q(\boldsymbol{z}) = \prod_{i=1}^{N} q_{z^{(i)}}(z^{(i)})$, where $i$ is an index over instances.
- The M-step updates typically correspond to the soft version of a probabilistic supervised learning algorithm, like Naïve Bayes.

Here are a few of the many applications of this general framework.

### 4.2.1 Word sense induction

The chapter began by considering the problem of word sense disambiguation when the senses are not known in advance. Expectation-maximization can be applied to this problem by treating each cluster as a word sense. Each instance represents the use of an ambiguous word, and $\boldsymbol{x}^{(i)}$ is a vector of counts for the other words that appear nearby: Schütze (1998) uses all words within a 50-word window. The probability $\text{p}(\boldsymbol{x}^{(i)} \mid y)$ can be set to the multinomial distribution, as in Naïve Bayes. The EM algorithm can be applied directly to this data, yielding clusters that (hopefully) correspond to the word senses.

Better performance can be obtained by first applying truncated **singular value decomposition (SVD)** to the matrix of context-counts $\mathbf{C}_{ij} = \text{count}(i, j)$, where $\text{count}(i, j)$ is the

count of word $j$ in the context of instance $i$.

$$\min_{\mathbf{U},\mathbf{S},\mathbf{V}} ||\mathbf{C} - \mathbf{U}\mathbf{S}\mathbf{V}^\top||_F \quad [4.25]$$
$$s.t. \mathbf{U} \in \mathbb{R}^{V \times K}, \mathbf{U}\mathbf{U}^\top = \mathbb{I}$$
$$\mathbf{S} = \text{Diag}(s_1, s_2, \ldots, s_K)$$
$$\mathbf{V}^\top \in \mathbb{R}^{N_p \times K}, \mathbf{V}\mathbf{V}^\top = \mathbb{I},$$

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 $\boldsymbol{c}_i$. EM clustering can be made more robust by setting the instance descriptions $\boldsymbol{x}^{(i)}$ equal to these rows, rather than using raw counts (Schütze, 1998). However, because the instances are now dense vectors of continuous numbers, the probability $\text{p}(\boldsymbol{x}^{(i)} \mid z)$ 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 $\mathbf{U}\mathbf{S}\mathbf{V}^\top$ 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 13.

### 4.2.2 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 ground truth annotations, ensuring that each label $y$ corresponds to the desired concept. At the same time, by making use of unlabeled data, it is possible cover a greater fraction of the features than would be possible using labeled data alone. Other methods for semi-supervised learning are discussed in § 4.3, but for now, we approach the problem within the framework of expectation-maximization (Nigam et al., 2000).

Suppose we have labeled data $\{\boldsymbol{x}^{(i)}, y^{(i)}\}_{i=1}^{N_\ell}$, and unlabeled data $\{\boldsymbol{x}^{(i)}\}_{i=N_\ell+1}^{N_\ell+N_u}$, where $N_\ell$ is the number of labeled instances and $N_u$ is the number of unlabeled instances. We can learn from the combined data by maximizing the joint log-likelihood,

$$\mathcal{L} = \sum_{i=1}^{N_\ell} \log \text{p}(\boldsymbol{x}^{(i)}, y^{(i)}; \boldsymbol{\mu}, \boldsymbol{\phi}) + \sum_{j=N_\ell+1}^{N_\ell+N_u} \log \text{p}(\boldsymbol{x}^{(j)}; \boldsymbol{\mu}, \boldsymbol{\phi}) \quad [4.26]$$
$$= \sum_{i=1}^{N_\ell} \Big( \log \text{p}(\boldsymbol{x}^{(i)} \mid y^{(i)}; \boldsymbol{\phi}) + \log \text{p}(y^{(i)}; \boldsymbol{\mu}) \Big) + \sum_{j=N_\ell+1}^{N_\ell+N_u} \log \sum_{y=1}^{K} \text{p}(\boldsymbol{x}^{(j)}, y; \boldsymbol{\mu}, \boldsymbol{\phi}). \quad [4.27]$$

The left sum is identical to the objective in Naïve Bayes; the right sum is the marginal log-likelihood for expectation-maximization clustering, from Equation 4.5. We can construct a lower bound on this log-likelihood by introducing distributions $\boldsymbol{q}^{(j)}$ for all $j \in \{N_\ell + 1, \ldots, N_\ell + N_u\}$. During the E-step, we update these distributions; during the M-step we update the parameters $\boldsymbol{\phi}$ and $\boldsymbol{\mu}$ using the expected counts from the unlabeled data and the observed counts from the labeled data.

A critical issue in semi-supervised learning is how to balance the impact of the labeled and unlabeled data on the classifier weights, especially when the unlabeled data is much larger than the labeled dataset. The risk is that the unlabeled data will dominate, causing the parameters to drift towards a "natural clustering" of the instances — which may not correspond to a good classifier for the labeled data. One solution is to heuristically reweight the two components of Equation 4.26, which can be critical to achieving good performance with EM (Nigam et al., 2000).

### 4.2.3 Multi-component modeling

As a final application, let's return to fully supervised classification. One of the classes in 20 newsgroups is `comp.sys.mac.hardware`; suppose that within this newsgroup there are two kinds of posts: reviews of new hardware, and question-answer posts about hardware problems. The language in these **components** of the `mac.hardware class` might have little in common; if so, it would be better to model these components separately, rather than treating their union as a single class. However, the component responsible for each instance is not directly observed.

Recall that Naïve Bayes is based on a generative process, which provides a stochastic explanation for the observed data. In Naïve Bayes, each label is drawn from a categorical distribution with parameter $\boldsymbol{\mu}$, and each vector of word counts is drawn from a multinomial distribution with parameter $\boldsymbol{\phi}_y$. For multi-component modeling, we envision a slightly different generative process, incorporating both the observed label $y^{(i)}$ and the latent component $z^{(i)}$:

- For each document $i$,
  - draw the label $y^{(i)} \sim \text{Categorical}(\boldsymbol{\mu})$
  - draw the component $z^{(i)} \mid y^{(i)} \sim \text{Categorical}(\boldsymbol{\beta}_{y^{(i)}})$, where $z^{(i)} \in 1, 2, \ldots, K_z$.
  - draw the vector of counts $\boldsymbol{x}^{(i)} \mid z^{(i)} \sim \text{Multinomial}(\boldsymbol{\phi}_{z^{(i)}})$

Our labeled data includes $\langle \boldsymbol{x}^{(i)}, y^{(i)} \rangle$, 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

| | | |
|---|---|---|
| (4.1) | ☺ | Villeneuve a bel et bien **réussi** son pari de changer de perspectives tout en assurant une cohérence à la franchise.[2] |
| (4.2) | ☹ | Il est également trop **long** et bancal dans sa narration, tiède dans ses intentions, et tiraillé entre deux personnages et directions qui ne parviennent pas à coexister en harmonie.[3] |
| (4.3) | | Denis Villeneuve a **réussi** une suite **parfaitement** maitrisée[4] |
| (4.4) | | **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] |
| (4.5) | | Une suite d'une écrasante puissance, mêlant **parfaitement** le contemplatif au narratif.[6] |
| (4.6) | | 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 décapités, l'armée des graphistes numériques griffant nerveusement la palette graphique entre agonie et orgasme.[7] |

Table 4.1: Labeled and unlabeled reviews of the films *Blade Runner 2049* and *Transformers: The Last Knight*.

lower bound on the log-likelihood,

$$\log \mathrm{p}(\boldsymbol{x}^{(i)}, y^{(i)}) = \log \sum_{z=1}^{K_z} \mathrm{p}(\boldsymbol{x}^{(i)}, y^{(i)}, z; \boldsymbol{\mu}, \boldsymbol{\phi}, \boldsymbol{\beta}) \tag{4.28}$$

$$\geq \log \mathrm{p}(y^{(i)}; \boldsymbol{\mu}) + E_q \left[ \log \mathrm{p}(\boldsymbol{x}^{(i)} \mid z; \boldsymbol{\phi}) + \log \mathrm{p}(z \mid y^{(i)}; \boldsymbol{\beta}) - \log q^{(i)}(z) \right]. \tag{4.29}$$

We are now ready to apply expectation maximization. As usual, E-step updates the distribution over the missing data, $q^{(i)}$. During the M-step, we update the parameters,

$$\beta_{y,z} = \frac{E_q\left[\text{count}(y, z)\right]}{\sum_{z'=1}^{K_z} E_q\left[\text{count}(y, z')\right]} \tag{4.30}$$

$$\phi_{z,j} = \frac{E_q\left[\text{count}(z, j)\right]}{\sum_{j'=1}^{V} E_q\left[\text{count}(z, j')\right]}. \tag{4.31}$$

## 4.3 Semi-supervised learning

In semi-supervised learning, the learner makes use of both labeled and unlabeled data. To see how this could help, suppose you want to do sentiment analysis in French. In Table 4.1, there are two labeled examples, one positive and one negative. From this data, a learner conclude that *réussi* is positive and *long* is negative; this isn't much! However, we can propagate this information to the unlabeled data, and potentially learn more.

- If we are confident that *réussi* is positive, then we might guess that (4.3) is also positive.
- That suggests that *parfaitement* is also positive.
- We can then propagate this information to (4.5), and learn more.
- Similarly, we can propagate from the labeled data to (4.4), which we guess to be negative because it shares the word *long*. This suggests that *bavard* is also negative, which we propagate to (4.6).

Instances (4.3) and (4.4) were "similar" to our labeled examples for positivity and negativity, respectively. By using these instances to expand the models for each class, it became possible to correctly label instances (4.5) and (4.6), which didn't share any important features with our original labeled data. This requires a key assumption: that similar instances will have similar labels.

In § 4.2.2, we saw how expectation-maximization can be applied to semi-supervised learning. In this case, you can imagine that the initial parameters $\phi$ assigned a high weight for *réussi* in the positive class, and a high weight for *long* in the negative class. These weights helped to shape the distributions $\boldsymbol{q}$ for instances (4.3) and (4.4) in the E-step. In the next iteration of the M-step, the parameters $\phi$ are updated with counts from these instances, making it possible to correctly label the instances (4.5) and (4.6).

However, expectation-maximization has a key disadvantage: it requires using a generative classification model, which imposes severe restrictions on the features that can be used for classification. In this section, we explore non-probabilistic approaches, which impose fewer restrictions on the classification model.

### 4.3.1 Multi-view learning

EM semi-supervised learning can be viewed as **self-training**: the labeled data guides the initial estimates of the classification parameters; these parameters are used to compute a label distribution over the unlabeled instances, $\boldsymbol{q}^{(i)}$; the label distributions are used to update the parameters. The risk is that self-training drifts away from the original labeled data. This problem can be ameliorated by multi-view learning. The key assumption is that the features can be decomposed into multiple "views", each of which is conditionally independent, given the label. For example, consider the problem of classifying a name as a person or location: one view is the name itself; another is the context in which it appears. This situation is illustrated in Table 4.2.

**Co-training** is an iterative multi-view learning algorithm, in which we have separate classifiers for each view (Blum and Mitchell, 1998). At each iteration of the algorithm, each classifier predicts labels for a subset of the unlabeled instances, using only the features available in its view. These predictions are then used as ground truth to train the classifiers

| | $\boldsymbol{x}^{(1)}$ | $\boldsymbol{x}^{(2)}$ | $y$ |
|---|---|---|---|
| 1. | Peachtree Street | located on | LOC |
| 2. | Dr. Walker | said | PER |
| 3. | Zanzibar | located in | ? $\rightarrow$ LOC |
| 4. | Zanzibar | flew to | ? $\rightarrow$ LOC |
| 5. | Dr. Robert | recommended | ? $\rightarrow$ PER |
| 6. | Oprah | recommended | ? $\rightarrow$ PER |

Table 4.2: Example of multiview learning for named entity classification

associated with the other views. In the example shown in Table 4.2, the classifier on $\boldsymbol{x}^{(1)}$ might correctly label instance #5 as a person, because of the feature *Dr*; this instance would then serve as training data for the classifier on $\boldsymbol{x}^{(2)}$, which would then be able to correctly label instance #6, thanks to the feature *recommended*. If the views are truly independent, this procedure is robust to drift. Furthermore, it imposes no restrictions on the classifiers that can be used for each view.

Word-sense disambiguation is particularly suited to multi-view learning, thanks to the heuristic of "one sense per discourse": if a polysemous word is used more than once in a given text or conversation, all usages refer to the same sense (Gale et al., 1992). This motivates a multi-view learning approach, in which one view corresponds to the local context, and another view corresponds to the global context at the document level (Yarowsky, 1995). We first train the local context view on a small seed dataset, and then identify its most confident predictions on unlabeled instances. The global context view is then used to extend these confident predictions to other instances within the same document. These new instances are added to the training data to the local context classifier, which is retrained and then re-applied to the remaining unlabeled data.

### 4.3.2 Graph-based algorithms

Another family of approaches to semi-supervised learning begins by constructing a graph, in which pairs of instances are linked with symmetric weights $w_{i,j}$, e.g.,

$$w_{i,j} = \exp(-\alpha \times ||\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}||^2). \quad [4.32]$$

Our goal is to use this weighted graph to propagate labels from a small set of labeled instances to larger set of unlabeled instances.

In **label propagation**, this is done through a series of matrix operations (Zhu et al., 2003). Let $\mathbf{Q}$ be a matrix of size $N \times K$, in which each row $\boldsymbol{q}^{(i)}$ describes the labeling of instance $i$. When ground truth labels are available, we set $\boldsymbol{q}^{(i)}$ to an indicator vector, with $q^{(i)}_{y^{(i)}} = 1$ and $q^{(i)}_{y' \neq y^{(i)}} = 0$. We will refer to the submatrix of rows containing labeled

instances as $\mathbf{Q}_L$, and the remaining rows as $\mathbf{Q}_U$. The rows of $\mathbf{Q}_U$ can be initialized to assign equal probabilities to all labels, $q_{i,k} = \frac{1}{K}$.

Now, let $T_{i,j}$ represent the "transition" probability of moving from node $j$ to node $i$,

$$T_{i,j} \triangleq \Pr(j \rightarrow i) = \frac{w_{i,j}}{\sum_{l=1}^{N} w_{k,j}}. \quad [4.33]$$

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 manipulate; a solution is to sparsify it, by keeping only the $\kappa$ largest values in each row, and setting all other values to zero. We can then "propagate" the label distributions to the unlabeled instances,

$$\tilde{\mathbf{Q}}_U \leftarrow \mathbf{T}\mathbf{Q} \quad [4.34]$$

$$\boldsymbol{s} \leftarrow \tilde{\mathbf{Q}}_U \mathbf{1} \quad [4.35]$$

$$\mathbf{Q}_U \leftarrow \text{Diag}(\boldsymbol{s})^{-1} \tilde{\mathbf{Q}}_U. \quad [4.36]$$

The second two lines normalize the rows of $\tilde{\mathbf{Q}}$, so that each row of $\mathbf{Q}$ is a probability distribution over labels.

## 4.4 Domain adaptation

We now consider a scenario in which the labeled data differs in some key respect from the unlabeled data to which we want to apply our trained model. A classic example is in consumer reviews: we may have labeled reviews of movies (the **source domain**), but we want to predict the reviews of appliances (the **target domain**). Similar issues arise in cases of genre: most linguistically-annotated data is in the domain of news text, but application domains include social media, electronic health records, and the text of government regulations. In general, there may be several source and target domains, each with their own properties; however, for simplicity we will focus mainly on the case of a single source and target domain.

The simplest approach is "direct transfer": train a classifier on the source domain, and apply it directly to the target domain. The accuracy of this approach will depend on the extent to which features are shared across domains. In review text, words like *outstanding* and *disappointing* will apply across both movies and appliances; but others, like *terrifying*, may have meanings that are domain-specific. **Domain adaptation** algorithms attempt to do better than direct transfer, by learning from data in both domains. There are two main families of domain adaptation algorithms, depending on whether any labeled data is available in the target domain.

### 4.4.1 Supervised domain adaptation

In supervised domain adaptation, we assume a small amount of labeled data in the target domain. The simplest approach would be to ignore domain differences, and simply merge the training data from the source and target domains. There are several other baseline approaches to dealing with this scenario (Daumé III, 2007):

- **Interpolation**: train a classifier for each domain, and combine their predictions. For example,
$$\hat{y} = \underset{y}{\text{argmax}}\, \lambda_s \boldsymbol{\theta}_s \cdot \boldsymbol{f}_s(\boldsymbol{x}, y) + (1 - \lambda_s)\boldsymbol{\theta}_t \cdot \boldsymbol{f}_t(\boldsymbol{x}, y), \quad [4.37]$$
where $\boldsymbol{\theta}_s$ and $\boldsymbol{\theta}_t$ are the weights of the source and target classifiers, and $\lambda_s$ is the interpolation weight on the source domain.
- **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.
- **Priors**: train a classifier on the source domain data, and use its weights as a prior distribution on the weights of the classifier for the target domain data. Equivalently, we can regularize the target domain weights towards the weights of the source domain classifier (Chelba and Acero, 2006),
$$\ell(\boldsymbol{\theta}_t) = \sum_{i=1}^{N} \ell^{(i)}(\boldsymbol{x}^{(i)}, y^{(i)}; \boldsymbol{\theta}_t) + \lambda ||\boldsymbol{\theta}_t - \boldsymbol{\theta}_s||_2^2, \quad [4.38]$$
where $\ell^{(i)}$ is the prediction loss on instance $i$, and $\lambda$ is the regularization weight.

An effective and "frustratingly simple" alternative is `EasyAdapt` (Daumé III, 2007). The main idea is to create 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,

$$\begin{aligned}\boldsymbol{f}(\boldsymbol{x}, y, d) = \{&\langle \textit{boring}, -, \text{MOVIE}\rangle, \langle \textit{boring}, -, *\rangle, \\ &\langle \textit{flavorless}, -, \text{MOVIE}\rangle, \langle \textit{flavorless}, -, *\rangle, \\ &\langle \textit{three-day-old}, -, \text{MOVIE}\rangle, \langle \textit{three-day-old}, -, *\rangle, \\ &\ldots\},\end{aligned}$$

with the asterisk indicating the cross-domain copy of each feature. It is then up to the learner to allocate weight between the domain-specific and cross-domain features: for

[8] `http://www.colesmithey.com/capsules/2017/06/wonder-woman.HTML`, accessed October 9. 2017.

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]

### 4.4.2 Unsupervised domain adaptation

In unsupervised domain adaptation, we have only unlabeled 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

Many unsupervised domain adaptation algorithms make use of linear projection,

$$\boldsymbol{f}(\boldsymbol{x}^{(i)}) = \mathbf{A}\boldsymbol{x}^{(i)}. \quad [4.39]$$

The projected vectors $\boldsymbol{f}(\boldsymbol{x}^{(i)})$ can then be used as base features during both training (from the source domain) and prediction (on the target domain).

The projection matrix $\mathbf{U}$ can be learned in a number of different ways, but many approaches focus on compressing and reconstructing the features in $\mathbf{X}$ (Ando and Zhang, 2005). For example, we can define a set of **pivot features**, which are typically chosen because they appear in both domains: in the case of review documents, pivot features might include evaluative adjectives like *outstanding* and *disappointing* (Blitzer et al., 2007). For each pivot feature $j$, we define an auxiliary problem of predicting whether the feature is present in each example, using the remaining base features. Let the weights of this classifier be written $\boldsymbol{\phi}_j$, and let us horizontally concatenate the weights for each of the $N_p$ pivot features into a matrix $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_{N_P}]$.

We then perform truncated singular value decomposition on $\boldsymbol{\Phi}$, as described in § 4.2.1, obtaining $\boldsymbol{\Phi} \approx \mathbf{U}\mathbf{S}\mathbf{V}^\top$. Now, the rows of the matrix $\mathbf{U}$ summarize information about each base feature: indeed, the truncated singular value decomposition identifies a low-dimension basis for the weight matrix $\Phi$, which in turn links base features to pivot features. Suppose that a base feature *reliable* occurs only in the target domain of appliance reviews. Nonetheless, it will have a positive weight towards some pivot features (e.g., *outstanding*, *recommended*), and a negative weight towards others (e.g., *worthless*, *unpleasant*). A base feature such as *watchable* might have the same associations with the pivot features,

[9] `EasyAdapt` 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).

and therefore, $\boldsymbol{u}_{\text{reliable}} \approx \boldsymbol{u}_{\text{watchable}}$. The matrix $\mathbf{U}$ can thus project the base features into a space in which this information is shared, $\boldsymbol{f}(\boldsymbol{x}^{(i)}) = \mathbf{U}^\top \boldsymbol{x}^{(i)}$.

#### Non-linear projection

More recent work has focused on learning non-linear transformations of the base features into a shared feature space that enables cross-domain transfer. This is accomplished by implementing the transformation function $\boldsymbol{f}$ as a deep neural network, which is trained from an auxiliary objective.

**Denoising objectives** One possibility is to train $\boldsymbol{f}$ to reconstruct a corrupted version of the original input. The original input can be corrupted in various ways: by the addition of random noise (Glorot et al., 2011; Chen et al., 2012), or by the deletion of features (Chen et al., 2012; Yang and Eisenstein, 2015). Denoising objectives share many properties of the linear projection method described above: they enable $\boldsymbol{f}$ to be trained on large amounts of unlabeled data from the target domain, and allow information to be shared across the feature space, thereby reducing sensitivity to rare and domain-specific features.

**Domain-generality objectives** Our ultimate goal is for the transformed representations $\boldsymbol{f}(\boldsymbol{x}^{(i)})$ to be domain-general. We can make this an explicit optimization criterion, by computing the similarity of transformed instances both within and between domains (Tzeng et al., 2015), or by formulating an additional classification task, in which the domain itself is treated as a label (Ganin et al., 2016). This setting is **adversarial**, because we want to learn a representation that makes this classifier perform poorly. At the same time, we want $\boldsymbol{f}(\boldsymbol{x}^{(i)})$ to enable accurate predictions of the labels $y^{(i)}$.

To formalize this idea, let $d^{(i)}$ represent the domain of instance $i$, and let $\ell(\boldsymbol{f}(\boldsymbol{x}^{(i)}), d^{(i)}; \boldsymbol{\theta}_d)$ represent the loss of a classifier (typically a deep neural network) trained to predict $d^{(i)}$ from the transformed representation $\boldsymbol{f}(\boldsymbol{x}^{(i)})$, using parameters $\boldsymbol{\theta}_d$. Analogously, let $\ell(\boldsymbol{f}(\boldsymbol{x}^{(i)}), y^{(i)}; \boldsymbol{\theta}_y)$ represent the loss of a classifier trained to predict the label $y^{(i)}$ from $\boldsymbol{f}(\boldsymbol{x}^{(i)})$, using parameters $\boldsymbol{\theta}_y$. The transformation $\boldsymbol{f}$ can then be trained from two criteria: we want it to yield accurate predictions of the labels $y^{(i)}$, while making **inaccurate** predictions of the domains $d^{(i)}$. This can be formulated as a joint optimization problem,

$$\min_{\boldsymbol{f}, \boldsymbol{\theta}_y, \boldsymbol{\theta}_d} \sum_{i=1}^{N_\ell + N_u} \ell(\boldsymbol{f}(\boldsymbol{x}^{(i)}), d^{(i)}; \boldsymbol{\theta}_d) - \sum_{i=1}^{N_\ell} \ell(g(\boldsymbol{x}^{(i)}), y^{(i)}; \boldsymbol{\theta}_y), \tag{4.40}$$

where $N_\ell$ is the number of labeled instances and $N_u$ is the number of unlabeled instances, with the labeled instances appearing first in the dataset. The loss can be optimized by stochastic gradient descent, jointly training the parameters of the non-linear transformation $\boldsymbol{f}$ and the parameters of the prediction models, $\boldsymbol{\theta}_d$ and $\boldsymbol{\theta}_y$.

## 4.5 *Other approaches to learning with latent variables

Expectation maximization is a very general way to think about learning with latent variables, but it has some limitations. One is the sensitivity to initialization, which means that we cannot simply run EM once and expect to get a good solution: in practical applications, considerable attention may need to be devoted to finding a good initialization. A second issue is that EM tends to be easiest to apply in cases where the latent variables have a clear decomposition (in the cases we have considered, they decompose across the instances). For these reasons, it is worth briefly considering some alternatives to EM.

### 4.5.1 Sampling

In EM clustering, we maintain a distribution $\boldsymbol{q}^{(i)}$ for the missing data related to each instance. In sampling-based algorithms, rather than maintaining a distribution over each latent variable, we draw random samples of the latent variables. If the sampling distribution is designed correctly, this procedure will eventually converge to drawing samples from the true posterior over the missing data, $\text{p}(\boldsymbol{z}^{(1:N_z)} \mid \boldsymbol{x}^{(1:N_x)})$. For example, in the case of clustering, the missing data $\boldsymbol{z}^{(1:N_z)}$ is the set of cluster memberships, $\boldsymbol{y}^{(1:N)}$, so we draw samples from the posterior distribution over clusterings of the data. If a single clustering is required, we can select the one with the highest conditional likelihood, $\hat{\boldsymbol{y}} = \text{argmax}_{\boldsymbol{y}} \, \text{p}(\boldsymbol{y}^{(1:N)} \mid \boldsymbol{x}^{(1:N)})$.

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,

$$z^{(n)} \mid \boldsymbol{x}, \boldsymbol{z}^{(-n)} \sim \text{p}(z^{(n)} \mid \boldsymbol{x}, \boldsymbol{z}^{(-n)}), \tag{4.41}$$

where $\boldsymbol{z}^{(-n)}$ indicates $\{\boldsymbol{z} \backslash z^{(n)}\}$, the set of all latent variables except for $z^{(n)}$. By repeatedly sampling over all latent variables, we construct a Markov chain that converges to a set of samples from the posterior, $\text{p}(\boldsymbol{z}^{(1:N_z)} \mid \boldsymbol{x}^{(1:N_x)})$. In probabilistic clustering, the sampling distribution has the following form,

$$\text{p}(y^{(i)} \mid \boldsymbol{x}, \boldsymbol{y}^{(-i)}) = \frac{\text{p}(\boldsymbol{x}^{(i)} \mid y^{(i)}; \boldsymbol{\phi}) \times \text{p}(y^{(i)}; \boldsymbol{\mu})}{\sum_{k=1}^{K} \text{p}(\boldsymbol{x}^{(i)} \mid k; \boldsymbol{\phi}) \times \text{p}(k; \boldsymbol{\mu})} \tag{4.42}$$

$$\propto \text{Multinomial}(\boldsymbol{x}^{(i)}; \boldsymbol{\phi}_{y^{(i)}}) \times \boldsymbol{\mu}_{y^{(i)}}. \tag{4.43}$$

Note that in this case, the sampling distribution does not depend on the other instances $\boldsymbol{x}^{(-i)}, \boldsymbol{y}^{(-i)}$: given the parameters $\boldsymbol{\phi}$ and $\boldsymbol{\mu}$, the posterior distribution over each $y^{(i)}$ can be computed from $\boldsymbol{x}^{(i)}$ alone.

In sampling algorithms, there are several choices for how to deal with the parameters. One possibility is to sample them too. To do this, we must add them to the generative story, by introducing a prior distribution. For the multinomial and categorical parameters in the EM clustering model, the **Dirichlet distribution** is a typical choice, since it defines a probability on exactly the set of vectors that can be parameters: vectors that sum to one and include only non-negative numbers.[10]

To incorporate this prior, we augment the generative model with the following lines,

- For each cluster $k \in \{1, 2, \ldots, K\}$,
  - Draw $\boldsymbol{\phi}_k \sim \text{Dirichlet}(\boldsymbol{\alpha}_\phi)$,
- Draw $\boldsymbol{\mu} \sim \text{Dirichlet}(\boldsymbol{\alpha}_\mu)$

where $\boldsymbol{\alpha}$ is a parameter of the prior distribution, typically set to a constant vector $\boldsymbol{\alpha} = [\alpha, \alpha, \ldots, \alpha]$. When $\alpha$ is large, the Dirichlet distribution tends to generate vectors that are nearly uniform; when $\alpha$ is small, it tends to generate vectors that assign most of their probability mass to a few entries. Given prior distributions over $\boldsymbol{\phi}$ and $\boldsymbol{\mu}$, we can now include them in Gibbs sampling, drawing values for these parameters from posterior distributions that are conditioned on the other variables in the model.

Unfortunately, sampling $\boldsymbol{\phi}$ and $\boldsymbol{\mu}$ can lead to slow convergence: the sampling distribution for these parameters is tightly constrained by the cluster memberships $\boldsymbol{y}^{(i)}$, which in turn is tightly constrained by the parameters. One alternative is to maintain $\boldsymbol{\phi}$ and $\boldsymbol{\mu}$ as parameters rather than latent variables. We can employ sampling in the E-step of the EM algorithm, constructing the distribution $\boldsymbol{q}^{(i)}$ from samples, and then updating the parameters using expected counts. This EM-MCMC hybrid is known as Monte Carlo Expectation Maximization (MCEM; Wei and Tanner, 1990), and is well-suited for cases in which it is difficult to compute $\boldsymbol{q}^{(i)}$ directly. Another possibility is to analytically integrate $\boldsymbol{\phi}$ and $\boldsymbol{\mu}$ out of the model. The cluster memberships $y^{(i)}$ are the only remaining latent variable; we sample them from the compound distribution,

$$\text{p}(y^{(i)} \mid \boldsymbol{x}^{(1:N)}, \boldsymbol{y}^{(-i)}; \alpha_\phi, \alpha_\mu) = \int_{\boldsymbol{\phi}, \boldsymbol{\mu}} \text{p}(\boldsymbol{\phi}, \boldsymbol{\mu} \mid \boldsymbol{y}^{(-i)}, \boldsymbol{x}^{(1:N)}; \alpha_\phi, \alpha_\mu) \text{p}(y^{(i)} \mid \boldsymbol{x}^{(1:N)}, \boldsymbol{y}^{(-i)}, \boldsymbol{\phi}, \boldsymbol{\mu}) d\boldsymbol{\phi} d\boldsymbol{\mu}. \quad [4.46]$$

[10] If $\sum_i^K \theta_i = 1$ and $\theta_i \geq 0$ for all $i$, then $\boldsymbol{\theta}$ is said to be on the $K-1$ **simplex**. A Dirichlet distribution with parameter $\boldsymbol{\alpha} \in \mathbb{R}_+^K$ has support over the $K-1$ simplex,

$$\text{p}_{\text{Dirichlet}}(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^K \theta_i^{\alpha_i - 1} \quad [4.44]$$

$$B(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^K \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^K \alpha_i)}, \quad [4.45]$$

with $\Gamma(\cdot)$ indicating the gamma function, a generalization of the factorial function to non-negative reals.

For multinomial and Dirichlet distributions, the sampling distribution can be computed in closed form. This is known as **collapsed Gibbs sampling**, because the parameters are "collapsed" out of the model.

MCMC algorithms are guaranteed to converge to the true posterior distribution over the latent variables, but there is no way to know how long this will take. In practice, the rate of convergence depends on initialization, just as expectation-maximization depends on initialization to avoid local optima. Thus, while Gibbs Sampling and other MCMC algorithms provide a powerful and flexible array of techniques for statistical inference in latent variable models, they are not a panacea for the problems experienced by EM.

### 4.5.2 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. Let us write $\overline{\boldsymbol{x}}^{(i)}$ for the normalized vector of word counts in document $i$, so that $\overline{\boldsymbol{x}}^{(i)} = \boldsymbol{x}^{(i)} / \sum_{j=1}^{V} x_j^{(i)}$. Then we can form a matrix of word-word co-occurrence counts,

$$\mathbf{C} = \sum_{i=1}^{N} \overline{\boldsymbol{x}}^{(i)} (\overline{\boldsymbol{x}}^{(i)})^\top. \tag{4.47}$$

We can also compute the expected value of this matrix under $\text{p}(\boldsymbol{x} \mid \boldsymbol{\phi}, \mu)$, as

$$E[\mathbf{C}] = \sum_{i=1}^{N} \sum_{k=1}^{K} \Pr(Z^{(i)} = k; \boldsymbol{\mu}) \boldsymbol{\phi}_k \boldsymbol{\phi}_k^\top \tag{4.48}$$

$$= \sum_{k}^{K} N \mu_k \boldsymbol{\phi}_k \boldsymbol{\phi}_k^\top \tag{4.49}$$

$$= \Phi \text{Diag}(N\mu) \Phi^\top, \tag{4.50}$$

where $\Phi$ is formed by horizontally concatenating $\boldsymbol{\phi}_1 \dots \boldsymbol{\phi}_K$, and $\text{Diag}(N\mu)$ indicates a diagonal matrix with values $N\mu_k$ at position $(k, k)$. Now, by setting $\mathbf{C}$ equal to its expectation, we obtain,

$$\mathbf{C} = \boldsymbol{\Phi} \text{Diag}(N\mu) \boldsymbol{\Phi}^\top, \tag{4.51}$$

which is very similar to the eigendecomposition $\mathbf{C} = \mathbf{Q} \boldsymbol{\Lambda} \mathbf{Q}^\top$. This suggests that simply by finding the eigenvectors and eigenvalues of $\mathbf{C}$, we could obtain the parameters $\boldsymbol{\phi}$ and $\mu$, and this is what motivates the name **spectral learning**.

While moment-matching and eigendecomposition are similar in form, there is a key difference in the constraints on the solutions: in eigendecomposition, we require orthonormality, so that $\mathbf{Q}\mathbf{Q}^\top = \mathbb{I}$; in estimating the parameters of a mixture model, we require that

$\boldsymbol{\mu}$ and the columns of $\Phi$ are probability vectors. Thus, spectral learning algorithms must include a procedure for converting the solution into vectors that are non-negative and sum to one. One approach is to replace eigendecomposition (or the related singular value decomposition) with non-negative matrix factorization (Xu et al., 2003), which guarantees that the solutions are non-negative (Arora et al., 2013).

After obtaining the parameters $\boldsymbol{\phi}$ and $\boldsymbol{\mu}$, we can obtain the distribution over clusters for each document by computing $\text{p}(z^{(i)} \mid \boldsymbol{x}^{(i)}; \boldsymbol{\phi}, \boldsymbol{\mu}) \propto \text{p}(\boldsymbol{x}^{(i)} \mid z^{(i)}; \boldsymbol{\phi})\text{p}(z^{(i)}; \mu)$. Spectral learning yields provably good solutions without regard to initialization, and can be quite fast in practice. However, it is more difficult to apply to a broad family of generative models than more generic techniques like EM and Gibbs Sampling. For more on applying spectral learning across a range of latent variable models, see Anandkumar et al. (2014).

## Additional reading

There are a number of other learning paradigms that deviate from supervised learning.

- **Active learning**: the learner selects unlabeled instances and requests annotations (Settles, 2012).
- **Multiple instance learning**: labels are applied to bags of instances, with a positive label applied if at least one instance in the bag meets the criterion (Dietterich et al., 1997; Maron and Lozano-Pérez, 1998).
- **Constraint-driven learning**: supervision is provided in the form of explicit constraints on the learner (Chang et al., 2007; Ganchev et al., 2010).
- **Distant supervision**: noisy labels are generated from an external resource (Mintz et al., 2009, also see § 16.2.3).
- **Multitask learning**: the learner induces a representation that can be used to solve multiple classification tasks (Collobert et al., 2011).
- **Transfer learning**: the learner must solve a classification task that differs from the labeled data (Pan and Yang, 2010).

Expectation maximization was introduced by Dempster et al. (1977), and is discussed in more detail by Murphy (2012). Like most machine learning treatments, Murphy focus on continuous observations and Gaussian likelihoods, rather than the discrete observations typically encountered in natural language processing. Murphy (2012) also includes an excellent chapter on MCMC; for an even more comprehensive treatment, see Robert and Casella (2013). For still more on Bayesian latent variable models, see Barber (2012), and for applications to natural language processing, see Cohen (2016). Surveys are available for semi-supervised learning (Zhu and Goldberg, 2009) and domain adaptation (Søgaard, 2013), although neither is recent enough to incorporate deep learning.

## Exercises

1. Derive the expectation maximization update for the parameter $\boldsymbol{\mu}$ in the EM clustering model.

2. The expectation maximization lower bound $\mathcal{J}$ is defined in Equation 4.10. Prove that the inverse $-\mathcal{J}$ is convex in $\boldsymbol{q}$. You can use the following facts about convexity:

   - $f(\boldsymbol{x})$ is convex in $\boldsymbol{x}$ iff $\alpha f(\boldsymbol{x}_1) + (1-\alpha) f(\boldsymbol{x}_2) \geq f(\alpha \boldsymbol{x}_1 + (1-\alpha)\boldsymbol{x}_2)$ for all $\alpha \in [0, 1]$.
   - If $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$ are both convex in $\boldsymbol{x}$, then $f(\boldsymbol{x}) + g(\boldsymbol{x})$ is also convex in $\boldsymbol{x}$.
   - $\log(x + y) \leq \log x + \log y$.

3. Derive the E-step and M-step updates for the following generative model. You may assume that the labels $y^{(i)}$ are observed, but $z_m^{(i)}$ is not.

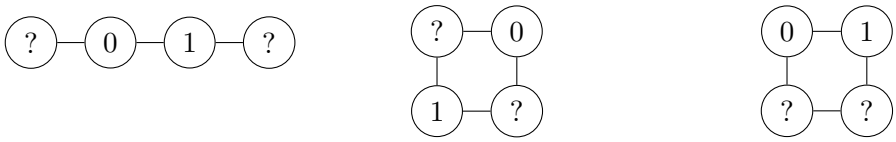
   - For each instance $i$,
     - Draw label $y^{(i)} \sim \text{Categorical}(\boldsymbol{\mu})$
     - For each token $m \in \{1, 2, \ldots, M^{(i)}\}$
       * Draw $z_m^{(i)} \sim \text{Categorical}(\pi)$
       * If $z_m^{(i)} = 0$, draw the current token from a label-specific distribution, $w_m^{(i)} \sim \boldsymbol{\phi}_{y^{(i)}}$
       * If $z_m^{(i)} = 1$, draw the current token from a document-specific distribution, $w_m^{(i)} \sim \boldsymbol{\nu}^{(i)}$

4. Use expectation-maximization clustering to train a word-sense induction system, applied to the word *say*.

   - Import `nltk`, run `nltk.download()` and select `semcor`. Import `semcor` from `nltk.corpus`.
   - The command `semcor.tagged_sentences(tag='sense')` returns an iterator over sense-tagged sentences in the corpus. Each sentence can be viewed as an iterator over `tree` objects. For `tree` objects that are sense-annotated words, you can access the annotation as `tree.label()`, and the word itself with `tree.leaves()`. So `semcor.tagged_sentences(tag='sense')[0][2].label()` would return the sense annotation of the third word in the first sentence.
   - Extract all sentences containing the senses `say.v.01` and `say.v.02`.
   - Build bag-of-words vectors $\boldsymbol{x}^{(i)}$, containing the counts of other words in those sentences, including all words that occur in at least two sentences.
   - Implement and run expectation-maximization clustering on the merged data.

- Compute the frequency with which each cluster includes instances of `say.v.01` and `say.v.02`.

5. Using the iterative updates in Equations 4.34-4.36, compute the outcome of the label propagation algorithm for the following examples.




The value inside the node indicates the label, $y^{(i)} \in \{0, 1\}$, with $y^{(i)} = ?$ for unlabeled nodes. The presence of an edge between two nodes indicates $w_{i,j} = 1$, and the absence of an edge indicates $w_{i,j} = 0$. For the third example, you need only compute the first three iterations, and then you can guess at the solution in the limit.

In the remaining exercises, you will try out some approaches for semisupervised learning and domain adaptation. You will need datasets in multiple domains. You can obtain product reviews in multiple domains here: `https://www.cs.jhu.edu/~mdredze/datasets/sentiment/processed_acl.tar.gz`. Choose a source and target domain, e.g. dvds and books, and divide the data for the target domain into training and test sets of equal size.

6. First, quantify the cost of cross-domain transfer.

   - Train a logistic regression classifier on the source domain training set, and evaluate it on the target domain test set.
   - Train a logistic regression classifier on the target domain training set, and evaluate it on the target domain test set. This it the "direct transfer" baseline.

   Compute the difference in accuracy, which is a measure of the transfer loss across domains.

7. Next, apply the **label propagation** algorithm from § 4.3.2.

   As a baseline, using only 5% of the target domain training set, train a classifier, and compute its accuracy on the target domain test set.

   Next, apply label propagation:

   - Compute the label matrix $\mathbf{Q}_L$ for the labeled data (5% of the target domain training set), with each row equal to an indicator vector for the label (positive or negative).

- Iterate through the target domain instances, including both test and training data. At each instance $i$, compute all $w_{ij}$, using Equation 4.32, with $\alpha = 0.01$. Use these values to fill in column $i$ of the transition matrix $\mathbf{T}$, setting all but the ten largest values to zero for each column $i$. Be sure to normalize the column so that the remaining values sum to one. You may need to use a sparse matrix for this to fit into memory.
- Apply the iterative updates from Equations 4.34-4.36 to compute the outcome of the label propagation algorithm for the unlabeled examples.

Select the test set instances from $\mathbf{Q}_U$, and compute the accuracy of this method. Compare with the supervised classifier trained only on the 5% sample of the target domain training set.

8. Using only 5% of the target domain training data (and all of the source domain training data), implement one of the supervised domain adaptation baselines in § 4.4.1. See if this improves on the "direct transfer" baseline from the previous problem

9. Implement `EasyAdapt` (§ 4.4.1), again using 5% of the target domain training data and all of the source domain data.

10. Now try unsupervised domain adaptation, using the "linear projection" method described in § 4.4.2. Specifically:

    - Identify 500 pivot features as the words with the highest frequency in the (complete) training data for the source and target domains. Specifically, let $x_i^d$ be the count of the word $i$ in domain $d$: choose the 500 words with the largest values of $\min(x_i^{\text{source}}, x_i^{\text{target}})$.
    - Train a classifier to predict each pivot feature from the remaining words in the document.
    - Arrange the features of these classifiers into a matrix $\mathbf{\Phi}$, and perform truncated singular value decomposition, with $k = 20$
    - Train a classifier from the source domain data, using the combined features $\boldsymbol{x}^{(i)} \oplus \mathbf{U}^\top \boldsymbol{x}^{(i)}$ — these include the original bag-of-words features, plus the projected features.
    - Apply this classifier to the target domain test set, and compute the accuracy.

# Part II

# Sequences and trees

# Chapter 5

# Language models

In probabilistic classification, we are interested in computing the probability of a label, conditioned on the text. Let us now consider something like the inverse problem: computing the probability of text itself. Specifically, we will consider models that assign probability to a sequence of word tokens,[1] $\text{p}(w_1, w_2, \ldots, w_M)$, with $w_m \in \mathcal{V}$. The set $\mathcal{V}$ is a discrete vocabulary,

$$\mathcal{V} = \{\textit{aardvark}, \textit{abacus}, \ldots, \textit{zither}\}. \quad [5.1]$$

Why would we want to compute the probability of a word sequence? In many applications, our goal is to produce word sequences as output:

- In **machine translation**, we convert from text in a source language to text in a target language.
- In **speech recognition**, we convert from audio signal to text.
- In **summarization**, we convert from long texts into short texts.
- In **dialogue systems**, we convert from the user's input (and perhaps an external knowledge base) into a text response.

In each of these cases, a key subcomponent is to compute the probability of the output text. By choosing high-probability output, we hope to generate texts that are more **fluent**. For example, suppose we want to translate a sentence from Spanish to English.

[1] The linguistic term "word" does not cover everything we might want to model, such as names, numbers, and emoticons. Instead, we prefer the term **token**, which refers to anything that can appear in a sequence of linguistic data. **Tokenizers** are programs for segmenting strings of characters or bytes into tokens. In standard written English, tokenization is relatively straightforward, and can be performed using a regular expression. But in languages like Chinese, tokens are not usually separated by spaces, so tokenization can be considerably more challenging. For more on tokenization algorithms, see Manning et al. (2008), chapter 2.

(5.1) *El cafe negro me gusta mucho*.

A literal word-for-word translation (sometimes called a **gloss**) is,

(5.2) *The coffee black me pleases much*.

A good language model of English will tell us that the probability of this translation is low, in comparison with more grammatical alternatives, such as,

$$\text{p}(\textit{The coffee black me pleases much}) < \text{p}(\textit{I love dark coffee}). \quad [5.2]$$

How can we use this fact? Warren Weaver, one of the early leaders in machine translation, viewed it as a problem of breaking a secret code (Weaver, 1955):

> When I look at an article in Russian, I say: 'This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.'

This observation motivates a generative model (like Naïve Bayes):

- The English sentence $\boldsymbol{w}^{(e)}$ is generated from a **language model**, $\text{p}_e(\boldsymbol{w}^{(e)})$.
- The Spanish sentence $\boldsymbol{w}^{(s)}$ is then generated from a **translation model**, $\text{p}_{s|e}(\boldsymbol{w}^{(s)} \mid \boldsymbol{w}^{(e)})$.

Given these two distributions, we can then perform translation by Bayes rule:

$$\text{p}_{e|s}(\boldsymbol{w}^{(e)} \mid \boldsymbol{w}^{(s)}) \propto \text{p}_{e,s}(\boldsymbol{w}^{(e)}, \boldsymbol{w}^{(s)}) \quad [5.3]$$

$$= \text{p}_{s|e}(\boldsymbol{w}^{(s)} \mid \boldsymbol{w}^{(e)}) \times \text{p}_e(\boldsymbol{w}^{(e)}). \quad [5.4]$$

This is sometimes called the **noisy channel model**, because it envisions English text turning into Spanish by passing through a noisy channel, $\text{p}_{s|e}$. What is the advantage of modeling translation this way, as opposed to modeling $\text{p}_{e|s}$ directly? The crucial point is that the two distributions $\text{p}_{s|e}$ (the translation model) and $\text{p}_e$ (the language model) can be estimated from separate data. The translation model requires **bitext** — examples of correct translations. But the language model requires only text in English. Such monolingual data is much more widely available, which means that the fluency of the output translation can be improved simply by scraping more webpages. Furthermore, once estimated, the language model $\text{p}_e$ can be reused in any application that involves generating English text, from summarization to speech recognition.

## 5.1 N-gram language models

How can we estimate the probability of a sequence of word tokens? The simplest idea would be to apply a **relative frequency estimator**. For example, consider the quote, attributed to Picasso, "*computers are useless, they can only give you answers.*" We can estimate the probability of this sentence as follows:

$$\begin{aligned}&\text{p}(\textit{Computers are useless, they can only give you answers}) \\ &= \frac{\text{count}(\textit{Computers are useless, they can only give you answers})}{\text{count}(\text{all sentences ever spoken})}\end{aligned} \quad [5.5]$$

This estimator is **unbiased**: in the theoretical limit of infinite data, the estimate will be correct. But in practice, we are asking for accurate counts over an infinite number of events, since sequences of words can be arbitrarily long. Even if we set an aggressive upper bound of, say, $n = 20$ tokens in the sequence, the number of possible sequences is $V^{20}$. A small vocabularly for English would have $V = 10^4$, so we would have $10^{80}$ possible sequences. Clearly, this estimator is very data-hungry, and suffers from high variance: even grammatical sentences will have probability zero if they happen not to have occurred in the training data.[2] We therefore need to introduce bias to have a chance of making reliable estimates from finite training data. The language models that follow in this chapter 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 $\text{p}(\boldsymbol{w}) = \text{p}(w_1, w_2, \ldots, w_M)$ can be refactored using the chain rule:

$$\text{p}(\boldsymbol{w}) = \text{p}(w_1, w_2, \ldots, w_M) \quad [5.6]$$

$$= \text{p}(w_1) \times \text{p}(w_2 \mid w_1) \times \text{p}(w_3 \mid w_2, w_1) \times \ldots \times \text{p}(w_M \mid w_{M-1}, \ldots, w_1) \quad [5.7]$$

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,

$$\begin{aligned}\text{p}(\textit{useless} \mid \textit{computers are}) &= \frac{\text{count}(\textit{computers are useless})}{\sum_{x \in \mathcal{V}} \text{count}(\textit{computers are } x)} \\ &= \frac{\text{count}(\textit{computers are useless})}{\text{count}(\textit{computers are})}.\end{aligned}$$

[2] Chomsky has 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*. Indeed, even the bigrams in these two examples are unlikely to occur — at least, not in texts written before Chomsky proposed this example.

Note that we haven't made any approximations yet, and we could have just as well applied the chain rule in reverse order,

$$\text{p}(\boldsymbol{w}) = \text{p}(w_M) \times \text{p}(w_{M-1} \mid w_M) \times \ldots \times \text{p}(w_1 \mid w_2, \ldots, w_M), \quad [5.8]$$

or in any other order. But this means that we also haven't really improved anything either: to compute the conditional probability $\text{p}(w_M \mid w_{M-1}, w_{M-2}, \ldots, w_1)$, we need to model $V^{M-1}$ contexts. We cannot estimate such a distribution from any reasonable finite sample.

To solve this problem, $n$-gram models make a crucial simplifying approximation: condition on only the past $n-1$ words.

$$\text{p}(w_m \mid w_{m-1} \ldots w_1) \approx \text{p}(w_m \mid w_{m-1}, \ldots, w_{m-n+1}) \quad [5.9]$$

This means that the probability of a sentence $\boldsymbol{w}$ can be computed as

$$\text{p}(w_1, \ldots, w_M) \approx \prod_m^M \text{p}(w_m \mid w_{m-1}, \ldots, w_{m-n+1}) \quad [5.10]$$

To compute the probability of an entire sentence, it is convenient to pad the beginning and end with special symbols $\lozenge$ and $\blacklozenge$. Then the bigram ($n = 2$) approximation to the probability of *I like black coffee* is:

$$\text{p}(\textit{I like black coffee}) = \text{p}(\textit{I} \mid \lozenge) \times \text{p}(\textit{like} \mid \textit{I}) \times \text{p}(\textit{black} \mid \textit{like}) \times \text{p}(\textit{coffee} \mid \textit{black}) \times \text{p}(\blacklozenge \mid \textit{coffee}). \quad [5.11]$$

In this model, we have to estimate and store the probability of only $V^n$ events, which is exponential in the order of the $n$-gram, and not $V^M$, which is exponential in the length of the sentence. The $n$-gram probabilities can be computed by relative frequency estimation,

$$\Pr(W_m = i \mid W_{m-1} = j, W_{m-2} = k) = \frac{\text{count}(i, j, k)}{\sum_{i'} \text{count}(i', j, k)} = \frac{\text{count}(i, j, k)}{\text{count}(j, k)} \quad [5.12]$$

A key design question is how to set the hyperparameter $n$, which controls the size of the context used in each conditional probability. If this is misspecified, the language model will sacrifice accuracy. Let's consider the potential problems concretely.

**When $n$ is too small.** Consider the following sentences:

(5.3) ***Gorillas*** *always like to groom* ***THEIR*** *friends.*

(5.4) *The* ***computer*** *that's on the 3rd floor of our office building* ***CRASHED****.*

The uppercase bolded words depend crucially on their predecessors in lowercase bold: the likelihood of *their* depends on knowing that *gorillas* is plural, and the likelihood of *crashed* depends on knowing that the subject is a *computer*. If the $n$-grams are not big enough to capture this context, then the resulting language model would offer probabilities that are too low for these sentences, and too high for sentences that fail basic linguistic tests like number agreement.

**When $n$ is too big.** In this case, we cannot make good estimates of the n-gram parameters from our dataset, because of data sparsity. To handle the *gorilla* example, we would need to model 6-grams; which means accounting for $V^6$ events. Under a very small vocabulary of $V = 10^4$, this means estimating the probability of $10^{24}$ distinct events.

These two problems point to another **bias-variance tradeoff**. A small $n$-gram size introduces high bias with respect to the true distribution, and a large $n$-gram size introduces high variance due to the huge number of possible events. But in reality the situation is even worse, because we often 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.

We will seek approaches 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**.

## 5.2 Smoothing and discounting

Limited data is a persistent problem in estimating language models. In § 5.1, we presented $n$-grams as a partial solution. But as we saw, 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 higher-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 (Chen and Goodman, 1999).

### 5.2.1 Smoothing

A major concern in language modeling is to avoid the situation $\text{p}(\boldsymbol{w}) = 0$, which could arise as a result of a single unseen n-gram. A similar problem arose in Naïve Bayes, and there we solved it by **smoothing**: adding imaginary "pseudo" counts. The same idea can be applied to $n$-gram language models, as shown here in the bigram case,

$$\text{p}_{\text{smooth}}(w_m \mid w_{m-1}) = \frac{\text{count}(w_{m-1}, w_m) + \alpha}{\sum_{w' \in \mathcal{V}} \text{count}(w_{m-1}, w') + V\alpha}. \quad [5.13]$$

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 $\alpha = 0.5$. Manning and Schütze (1999) offer more insight on the justifications for this setting.

To maintain normalization, anything that we add to the numerator ($\alpha$) must also appear in the denominator ($V\alpha$). This idea is reflected in the concept of **effective counts**:

$$c_i^* = (c_i + \alpha)\frac{M}{M + V\alpha}, \quad [5.14]$$

where $c_i$ is the count of event $i$, $c_i^*$ is the effective count, and $M = \sum_i^V c_i$ is the total number of terms in the dataset $(w_1, w_2, \ldots, w_M)$. This term ensures that $\sum_i^V c_i^* = \sum_i^V c_i = M$. The **discount** for each n-gram is then computed as,

$$d_i = \frac{c_i^*}{c_i} = \frac{(c_i + \alpha)}{c_i}\frac{M}{(M + V\alpha)}.$$

### 5.2.2 Discounting and backoff

Discounting "borrows" probability mass from observed n-grams and redistributes it. In Lidstone smoothing, we borrow probability mass by increasing the denominator of the relative frequency estimates, and redistribute it by increasing the numerator for all n-grams. But instead, we could borrow the same amount of probability mass from all observed counts, and redistribute it among only the unobserved counts. This is called **absolute discounting.** For example, suppose we set an absolute discount $d = 0.1$ in a bigram model, and then redistribute this probability mass equally over the unseen words. The resulting probabilities are shown in Table 5.1.

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. In other words, 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 this smoothing model, bigram probabilities are computed as,

$$c^*(i, j) = c(i, j) - d \quad [5.15]$$

$$\mathrm{p}_{\text{Katz}}(i \mid j) = \begin{cases} \frac{c^*(i,j)}{c(j)} & \text{if } c(i, j) > 0 \\ \alpha(j) \times \frac{\mathrm{p}_{\text{unigram}}(i)}{\sum_{i':c(i',j)=0} \mathrm{p}_{\text{unigram}}(i')} & \text{if } c(i, j) = 0. \end{cases} \quad [5.16]$$

The term $\alpha(j)$ indicates the amount of probability mass that has been discounted for context $j$. This probability mass is then divided across all the unseen events, $\{i' : c(i', j) =$

| | counts | unsmoothed probability | Lidstone smoothing, $\alpha = 0.1$ | | Discounting, $d = 0.1$ | |
|---|---|---|---|---|---|---|
| | | | effective counts | smoothed probability | effective counts | smoothed probability |
| *impropriety* | 8 | 0.4 | 7.826 | 0.391 | 7.9 | 0.395 |
| *offense* | 5 | 0.25 | 4.928 | 0.246 | 4.9 | 0.245 |
| *damage* | 4 | 0.2 | 3.961 | 0.198 | 3.9 | 0.195 |
| *deficiencies* | 2 | 0.1 | 2.029 | 0.101 | 1.9 | 0.095 |
| *outbreak* | 1 | 0.05 | 1.063 | 0.053 | 0.9 | 0.045 |
| *infirmity* | 0 | 0 | 0.097 | 0.005 | 0.25 | 0.013 |
| *cephalopods* | 0 | 0 | 0.097 | 0.005 | 0.25 | 0.013 |

Table 5.1: Example of Lidstone smoothing and absolute discounting in a bigram language model, for the context $(\textit{alleged}, \_)$, 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*.

$0\}$, proportional to the unigram probability of each word $i'$. The discount parameter $d$ can be optimized to maximize performance (typically held-out log-likelihood) on a development set.

### 5.2.3 *Interpolation

Backoff is one way to combine $n$-gram models across various values of $n$. An alternative approach is **interpolation**: setting the probability of a word in context to a weighted sum of its 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,

$$\begin{aligned} \mathrm{p}_{\text{Interpolation}}(i \mid j, k) = &\lambda_3 \mathrm{p}_3^*(i \mid j, k) \\ &+ \lambda_2 \mathrm{p}_2^*(i \mid j) \\ &+ \lambda_1 \mathrm{p}_1^*(i). \end{aligned}$$

In this equation, $\mathrm{p}_n^*$ is the unsmoothed empirical probability given by an $n$-gram language model, and $\lambda_n$ is the weight assigned to this model. To ensure that the interpolated $\mathrm{p}(\boldsymbol{w})$ is still a valid probability distribution, we must obey the constraint, $\sum_n \lambda_n = 1$. But how to find the specific values of $\lambda$?

An elegant solution is **expectation maximization**. Recall from chapter 4 that we can think about EM as learning with **missing data**: we just need to choose missing data such

that learning would be easy if it weren't missing. What's missing in this case? We can think of each word $w_m$ as drawn from an n-gram of unknown size, $z_m \in \{1 \dots n_{\text{max}}\}$. This $z_m$ is the missing data that we are looking for. Therefore, the application of EM to this problem involves the following **generative process**:

- For each token $m \in \{1, 2, \dots, M\}$:
  - draw $z_m \sim \text{Categorical}(\lambda)$,
  - draw $w_m \sim \text{p}^*_{z_m}(w_m \mid w_{m-1}, \dots w_{m-z_m})$.

If the missing data $\{Z_m\}$ were known, then we could estimate $\lambda$ from relative frequency estimation,

$$\lambda_z = \frac{\text{count}(Z_m = z)}{M} \tag{5.17}$$

$$\propto \sum_{m=1}^{M} \delta(Z_m = z). \tag{5.18}$$

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 $z_m$,

$$q_m(z) \triangleq \Pr(Z_m = z \mid \boldsymbol{w}_{1:m}; \lambda) \tag{5.19}$$

$$= \frac{\text{p}(w_m \mid \boldsymbol{w}_{1:m-1}, Z_m = z) \times \text{p}(z)}{\sum_{z'} \text{p}(w_m \mid \boldsymbol{w}_{1:m-1}, Z_m = z') \times \text{p}(z')} \tag{5.20}$$

$$\propto \text{p}^*_z(w_m \mid \boldsymbol{w}_{1:m-1}) \times \lambda_z. \tag{5.21}$$

In the M-step, we can compute $\lambda$ by summing the expected counts under $q$,

$$\lambda_z \propto \sum_{m=1}^{M} q_m(z). \tag{5.22}$$

By iterating between updates to $q$ and $\lambda$, we will ultimately converge at a solution. The complete algorithm is shown in Algorithm 9.

### 5.2.4 *Kneser-Ney smoothing

Kneser-Ney smoothing is based on absolute discounting, but it redistributes the resulting probability mass in a different way from Katz backoff. Empirical evidence points to Kneser-Ney smoothing as the state-of-art for $n$-gram language modeling (Goodman, 2001). To motivate Kneser-Ney smoothing, consider the example: *I recently visited* _. Which of the following is more likely?

**Algorithm 9** Expectation-maximization for interpolated language modeling

**procedure** ESTIMATE INTERPOLATED $n$-GRAM ($\boldsymbol{w}_{1:M}$, $\{\text{p}_n^*\}_{n \in 1:n_{\text{max}}}$)
  **for** $z \in \{1, 2, \ldots, n_{\text{max}}\}$ **do** ▷ Initialization
    $\lambda_z \leftarrow \frac{1}{n_{\text{max}}}$
  **repeat**
    **for** $m \in \{1, 2, \ldots, M\}$ **do** ▷ E-step
      **for** $z \in \{1, 2, \ldots, n_{\text{max}}\}$ **do**
        $q_m(z) \leftarrow \text{p}_z^*(w_m \mid \boldsymbol{w}_{1:m-}) \times \lambda_z$
      $\boldsymbol{q}_m \leftarrow \text{Normalize}(\boldsymbol{q}_m)$
    **for** $z \in \{1, 2, \ldots, n_{\text{max}}\}$ **do** ▷ M-step
      $\lambda_z \leftarrow \frac{1}{M} \sum_{m=1}^{M} q_m(z)$
  **until** tired
  **return** $\boldsymbol{\lambda}$

- *Francisco*
- *Duluth*

Now suppose that both bigrams *visited Duluth* and *visited Francisco* are unobserved in our training data, and furthermore, that the unigram probability $\text{p}_1^*(\textit{Francisco})$ is greater than $\text{p}^*(\textit{Duluth})$. Nonetheless we would still guess that $\text{p}(\textit{visited Duluth}) > \text{p}(\textit{visited Francisco})$, because *Duluth* is a more **versatile** word: it an occur in many contexts, while *Francisco* usually occurs in a single context, following the word *San*. This notion of versatility is the 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

$$\text{p}_{KN}(w \mid u) = \begin{cases} \frac{\text{count}(w,u) - d}{\text{count}(u)}, & \text{count}(w, u) > 0 \\ \alpha(u) \times \text{p}_{\text{continuation}}(w), & \text{otherwise} \end{cases} \quad [5.23]$$

$$\text{p}_{\text{continuation}}(w) = \frac{|u : \text{count}(w, u) > 0|}{\sum_{w' \in \mathcal{V}} |u' : \text{count}(w', u') > 0|}. \quad [5.24]$$

First, note that we reserve probability mass using absolute discounting $d$, which is taken from all unobserved $n$-grams. The total amount of discounting in context $u$ is $d \times |w : \text{count}(w, u) > 0|$, and we divide this probability mass equally among the unseen $n$-grams,

$$\alpha(u) = |w : \text{count}(w, u) > 0| \times \frac{d}{\text{count}(u)}. \quad [5.25]$$

This is the amount of probability mass left to account for versatility, which we define via the *continuation probability* $\text{p}_{\text{continuation}}(w)$ as proportional to the number of observed contexts in which $w$ appears. In the numerator of the continuation probability we have the number of contexts $u$ in which $w$ appears, and in the denominator, we normalize by summing the same quantity over all words $w'$.

The idea of modeling versatility by counting contexts may seem heuristic, but there is an elegant theoretical justification from Bayesian nonparametrics (Teh, 2006). Kneser-Ney smoothing on $n$-grams was the dominant language modeling technique — widely used in speech recognition and machine translation — before the arrival of neural language models.

## 5.3 Recurrent neural network language models

Until this decade, $n$-grams were the dominant language modeling approach. But in a few years, they have been almost completely supplanted by a new family of language models based on **neural networks**. These models do not make the $n$-gram assumption of restricted context; indeed, they can incorporate arbitrarily distant contextual information, while remaining computationally and statistically tractable.

The first insight is to treat word prediction as a **discriminative** learning task: rather than directly estimating the distribution $\text{p}(w \mid u)$ from (smoothed) relative frequencies, we now treat language modeling as a machine learning problem, and estimate parameters that maximize the log conditional probability of a corpus.[3]

The second insight is to reparametrize the probability distribution $\text{p}(w \mid u)$ as a function of two dense $K$-dimensional numerical vectors, $\boldsymbol{\beta}_w \in \mathbb{R}^K$, and $\boldsymbol{v}_u \in \mathbb{R}^K$,

$$\text{p}(w \mid u) = \frac{\exp(\boldsymbol{\beta}_w \cdot \boldsymbol{v}_u)}{\sum_{w' \in \mathcal{V}} \exp(\boldsymbol{\beta}_{w'} \cdot \boldsymbol{v}_u)}, \quad [5.26]$$

where $\boldsymbol{\beta}_w \cdot \boldsymbol{v}_u$ represents a dot product. Note that the denominator ensures that it is a properly normalized probability distribution. In the neural networks literature, this function is sometimes known as a **softmax** layer, written

$$(\text{SoftMax}(\boldsymbol{a}))_i = \frac{\exp(a_i)}{\sum_j \exp(a_j)}, \quad [5.27]$$

where $\boldsymbol{a}$ is a vector of scores and $\text{SoftMax}(\boldsymbol{a})_i$ is a normalized distribution.[4]

[3] This idea is not in itself new; for example, Rosenfeld (1996) applies logistic regression to language modeling, and Roark et al. (2007) apply perceptrons and conditional random fields (§ 6.5.3).

[4] The logistic regression classifier can be viewed as an application of the softmax transformation to the vector constructed by computing the inner products of weights and features for all possible labels.

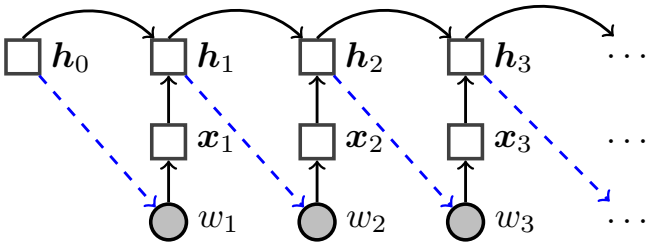



Figure 5.1: The recurrent neural network language model, viewed as an "unrolled" computation graph. Solid lines indicate direct computation, dotted blue lines indicate probabilistic dependencies, circles indicate random variables, and squares indicate computation nodes.

The word vectors $\boldsymbol{\beta}_w$ are parameters of the model, and are estimated directly. As we will see in chapter 13, these vectors carry useful information about word meaning, and semantically similar words tend to have highly correlated vectors.

The context vectors $\boldsymbol{v}_u$ can be computed in various ways, depending on the model. Here we will consider a relatively simple — but effective — neural language model, the **recurrent neural network** (RNN; Mikolov et al., 2010). The basic idea is to recurrently update the context vectors as we move through the sequence. Let us write $\boldsymbol{h}_m$ for the contextual information at position $m$ in the sequence. RNNs employ the following recurrence:

$$\boldsymbol{x}_m \triangleq \boldsymbol{\phi}_{w_m} \quad [5.28]$$

$$\boldsymbol{h}_m = g(\Theta \boldsymbol{h}_{m-1} + \boldsymbol{x}_m) \quad [5.29]$$

$$\text{p}(w_{m+1} \mid w_1, w_2, \ldots, w_m) = \frac{\exp(\boldsymbol{\beta}_{w_{m+1}} \cdot \boldsymbol{h}_m)}{\sum_{w' \in \mathcal{V}} \exp(\boldsymbol{\beta}_{w'} \cdot \boldsymbol{h}_m)}, \quad [5.30]$$

where $\boldsymbol{\phi}$ is a matrix of **input word embeddings**, and $\boldsymbol{x}_m$ denotes the embedding for word $w_m$. The function $g$ is an element-wise nonlinear **activation function**, as described in § 2.1.

A key point about the RNN language model is that although each $w_m$ depends only on the context vector $\boldsymbol{h}_{m-1}$, this vector is in turn influenced by *all* previous tokens, $w_1, w_2, \ldots w_{m-1}$, through the recurrence operation: $w_1$ affects $\boldsymbol{h}_1$, which affects $\boldsymbol{h}_2$, and so on, until the information is propagated all the way to $\boldsymbol{h}_{m-1}$, and then on to $w_m$ (see Figure 5.1). This is an important distinction from $n$-gram language models, where any information outside the $n$-word window is ignored. Thus, in principle, the RNN language model can handle long-range dependencies, such as number agreement over long spans of text — although it would be difficult to know where exactly in the vector $\boldsymbol{h}_m$ this information is represented. The main limitation is that information is attenuated by repeated application of the nonlinearity $g$. **Long short-term memories** (LSTMs), described below, are a variant of

RNNs that address this issue, using memory cells to propagate information through the sequence without applying non-linearities (Hochreiter and Schmidhuber, 1997).

The denominator in Equation 5.30 is a computational bottleneck, because it involves a sum over the entire vocabulary. One solution is to use a **hierarchical softmax** function, which computes the sum more efficiently by organizing the vocabulary into a tree (Mikolov et al., 2011). Another strategy is to optimize an alternative metric, such as **noise-contrastive estimation** (Gutmann and Hyvärinen, 2012), which learns by distinguishing observed instances from artificial instances generated from a noise distribution (Mnih and Teh, 2012).

### 5.3.1 Estimation by backpropagation

The recurrent neural network language model has the following parameters:

- $\boldsymbol{\phi}_i \in \mathbb{R}^K$, the "input" word vectors (these are sometimes called **word embeddings**, since each word is embedded in a $K$-dimensional space);
- $\boldsymbol{\beta}_i \in \mathbb{R}^K$, the "output" word vectors;
- $\Theta \in \mathbb{R}^{K \times K}$, the recurrence operator.

Each of these parameters must be estimated. We do this by formulating an objective function over the training corpus, $L(\boldsymbol{w})$, and then employ **backpropagation** to incrementally update the parameters after encountering each training example. Backpropagation is a term from the neural network literature, which means that we use the chain rule of differentiation to obtain gradients on each parameter. After obtaining these gradients, we can apply an online learning algorithm such as stochastic gradient descent (see § 1.5.2).

For example, suppose we want to obtain the gradient of the log-likelihood with respect to a single row of the recurrence operator, $\boldsymbol{\theta}_k$. Let us first define the total objective as a sum over local error functions $e_m$,

$$\ell(\boldsymbol{w}) = \sum_{m=1}^{M} e_m(\boldsymbol{h}_{m-1}) \quad [5.31]$$

$$e_m(\boldsymbol{h}_{m-1}) \triangleq -\log \mathrm{p}(w_m \mid w_1, w_2, \ldots, w_{m-1}) \quad [5.32]$$

$$= -\boldsymbol{\beta}_{w_m} \cdot \boldsymbol{h}_{m-1} + \log \sum_{w' \in \mathcal{V}} \exp\left(\boldsymbol{\beta}_{w'} \cdot \boldsymbol{h}_{m-1}\right). \quad [5.33]$$

We can now differentiate the objective with respect to $\boldsymbol{\theta}_k$:

$$\frac{\partial}{\partial \boldsymbol{\theta}_k}\ell(\boldsymbol{w}) = \sum_m \frac{\partial e_m(\boldsymbol{h}_{m-1})}{\partial \boldsymbol{\theta}_k} \quad [5.34]$$

$$= \sum_{m=1}^{M} (\nabla_{\boldsymbol{h}_{m-1}} e_m) \frac{\partial}{\partial \boldsymbol{\theta}_k} \boldsymbol{h}_{m-1}. \quad [5.35]$$

In the first line, we simply distribute the derivative across the sum. In the second line, we apply the chain rule of calculus. The term $\nabla_{\boldsymbol{h}_{m-1}} e_m$ refers to the gradient of the error $e_m$ evaluated at $\boldsymbol{h}_{m-1}$, and is equal to,

$$\nabla_{\boldsymbol{h}_{m-1}} e_m = -\boldsymbol{\beta}_{w_m} + B\,\text{SoftMax}(B\boldsymbol{h}_{m-1}), \quad [5.36]$$

where $B$ is a matrix with all word output embeddings stacked vertically, $B = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top, \ldots, \boldsymbol{\beta}_V^\top)$.

Next we compute the derivative of $\boldsymbol{h}_{m-1}$, first noting that within the vector $\boldsymbol{h}_{m-1}$, only the element $h_{m-1,k}$ depends on $\boldsymbol{\theta}_k$.

$$\frac{\partial}{\partial \boldsymbol{\theta}_k}\boldsymbol{h}_{m-1} = \frac{\partial}{\partial \boldsymbol{\theta}_k} h_{m-1,k} \quad [5.37]$$

$$= \frac{\partial}{\partial \boldsymbol{\theta}_k} g(\boldsymbol{\theta}_k \cdot \boldsymbol{h}_{m-2} + x_{m-1,k}) \quad [5.38]$$

$$= (\nabla_{\boldsymbol{\theta}_k \cdot \boldsymbol{h}_{m-2} + x_{m-1,k}} g) \times \frac{\partial}{\partial \boldsymbol{\theta}_k}(\boldsymbol{\theta}_k \cdot \boldsymbol{h}_{m-2} + x_{m-1,k}) \quad [5.39]$$

$$= (\nabla_{\boldsymbol{\theta}_k \cdot \boldsymbol{h}_{m-2} + x_{m-1,k}} g) \times (\boldsymbol{h}_{m-2} + \boldsymbol{\theta}_k \odot \frac{\partial}{\partial \boldsymbol{\theta}_k}\boldsymbol{h}_{m-2}), \quad [5.40]$$

where $\odot$ is an elementwise (Hadamard) vector product, and $(\nabla_{\boldsymbol{\theta}_k \cdot \boldsymbol{h}_{m-2} + x_{m-1,k}} g)$ is the elementwise gradient of the non-linear activation function for $\boldsymbol{h}_{m-1}$ evaluated at the scalar $\boldsymbol{\theta}_k \cdot \boldsymbol{h}_{m-2} + x_{m-1,k}$. For example, if $g$ is the elementwise hyperbolic tangent, then its gradient is,

$$(\nabla_{\boldsymbol{\theta}_k \cdot \boldsymbol{h}_{m-2} + x_{m-1,k}} g) = (1 - \tanh^2(\boldsymbol{\theta}_k \cdot \boldsymbol{h}_{m-2} + x_{m-1,k})). \quad [5.41]$$

The application of backpropagation to sequence models such as recurrent neural networks is known as **backpropagation through time**. A key point is that the derivative $\frac{\partial \boldsymbol{h}_{m-1}}{\partial \boldsymbol{\theta}_k}$ depends recurrently on $\frac{\partial \boldsymbol{h}_{m-2}}{\partial \boldsymbol{\theta}_k}$, and on all $\frac{\partial \boldsymbol{h}_n}{\partial \boldsymbol{\theta}_k}$ for $n < m$. Furthermore, we will need to compute $\frac{\partial \boldsymbol{h}_{m-2}}{\partial \boldsymbol{\theta}_k}$ **again**, to account for the error term $e_{m-1}(\boldsymbol{h}_{m-2})$. To avoid redoing work, it is best to cache such derivatives, so that they can be reused during backpropagation.

Backpropagation is implemented by neural network toolkits such as TensorFlow (Abadi et al., 2016), Torch (Collobert et al., 2011), and DyNet (Neubig et al., 2017). In these toolkits,

the user defines a **computation graph** representing the neural network structure, which culminates in a scalar loss function. The toolkit then automatically computes the gradient of the loss function with respect to all model parameters, by applying the chain rule of differentiation across the computation graph. Unlike the classification objectives considered in § 1.2, neural network objectives are usually non-convex function of the parameters, so there is no learning procedure that is guaranteed to converge to the global optimum. Nonetheless, gradient-based optimization often yields parameter estimates that are very effective in practice.

### 5.3.2 Hyperparameters

The RNN language model has several hyperparameters that must be tuned to ensure good performance. The model capacity is controlled by the size of the word and context vectors $K$, which play a role that is somewhat analogous to the size of the $n$-gram context. For datasets that are large with respect to the vocabulary (i.e., there is a large token-to-type ratio), we can afford to estimate a model with a large $K$, which enables more subtle distinctions between words and contexts. When the dataset is relatively small, then $K$ must be smaller too. However, this general advice has not yet been formalized into any concrete formula for choosing $K$, and trial-and-error is still necessary. Overfitting can also be prevented by **dropout**, which involves randomly setting some elements of the computation to zero (Srivastava et al., 2014), forcing the learner not to rely too much on any particular dimension of the word or context vectors. (The dropout rate must also be tuned by the user.) Other design decisions include: the nature of the nonlinear activation function $g$, the size of the vocabulary, and the parametrization of the learning algorithm, such as the learning rate.

### 5.3.3 Alternative neural language models

A well known problem with RNNs is that backpropagation across long chains tends to lead to "vanishing" or "exploding" gradients (Bengio et al., 1994). For example, the input embedding of word $w_1$ affects the likelihood of a distant word such as $w_{29}$, but this impact may be attenuated by backpropagation through the intervening time steps. One solution is to rescale the gradients, or to clip them at some maximum value (Pascanu et al., 2013). An alternative is to change the model architecture itself.

A popular variant of RNNs, which is more robust to these problems, is the **long short-term memory** (**LSTM**; Hochreiter and Schmidhuber, 1997; Sundermeyer et al., 2012). This model augments the hidden state $\boldsymbol{h}_m$ with a "memory cell" $\boldsymbol{c}_m$. The value of the memory cell at each time $m$ is a linear interpolation between two quantities: its previous value $\boldsymbol{c}_{m-1}$, and an "update" $\tilde{c}_m$, which is computed from the current input $\boldsymbol{x}_m$ and the previous hidden state $\boldsymbol{h}_{m-1}$. The next state $\boldsymbol{h}_m$ is then computed from the memory cell.

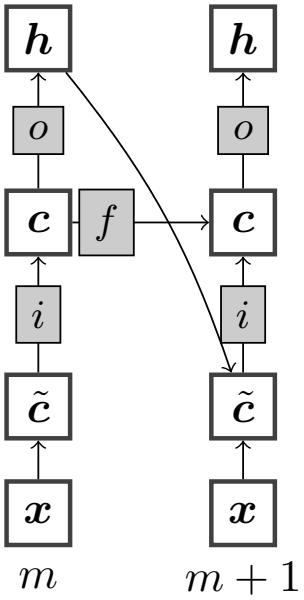



Figure 5.2: The long short-term memory (LSTM) architecture. In an LSTM language model, each $\boldsymbol{h}_m$ would be used to predict the next word $w_{m+1}$. [todo: redo with new computation graph format]

Because the memory cell is never passed through the non-linear function $g$, it is possible for information to propagate through the network over long distances.

The interpolation weights are controlled by a set of gates, which are themselves functions of the input and previous hidden state. The gates are computed from sigmoid activations, 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 5.2, and the complete update equations are:

$$\boldsymbol{f}_m = \sigma(\Theta^{(h\rightarrow f)} \cdot \boldsymbol{h}_{m-1} + \Theta^{(x\rightarrow f)} \cdot \boldsymbol{x}_m) \quad \text{forget gate} \quad [5.42]$$

$$\boldsymbol{i}_m = \sigma(\Theta^{(h\rightarrow i)} \cdot \boldsymbol{h}_{m-1} + \Theta^{(x\rightarrow i)} \cdot \boldsymbol{x}_m) \quad \text{input gate} \quad [5.43]$$

$$\tilde{\boldsymbol{c}}_m = \tanh(\Theta^{(h\rightarrow c)} \cdot \boldsymbol{h}_{m-1} + \Theta^{(w\rightarrow c)} \cdot \boldsymbol{x}_m) \quad \text{update candidate} \quad [5.44]$$

$$\boldsymbol{c}_m = \boldsymbol{f}_m \odot \boldsymbol{c}_{m-1} + \boldsymbol{i}_m \odot \tilde{\boldsymbol{c}}_m \quad \text{memory cell update} \quad [5.45]$$

$$\boldsymbol{o}_m = \sigma(\Theta^{(h\rightarrow o)} \cdot \boldsymbol{h}_{m-1} + \Theta^{(x\rightarrow o)} \cdot \boldsymbol{x}_m) \quad \text{output gate} \quad [5.46]$$

$$\boldsymbol{h}_m = \boldsymbol{o}_m \odot \boldsymbol{c}_m \quad \text{output.} \quad [5.47]$$

As above, $\odot$ refers to an elementwise (Hadamard) product. The LSTM model has been shown to outperform standard recurrent neural networks across a wide range of problems

(it was first used for language modeling by Sundermeyer et al. (2012)), and is now widely used for sequence modeling tasks. There are several LSTM variants, of which the Gated Recurrent Unit (Cho et al., 2014) is presently one of the more well known. Many software packages implement a variety of RNN architectures, so choosing between them is simple from a user's perspective. Jozefowicz et al. (2015) provide an empirical comparison of various modeling choices circa 2015. Notable earlier non-recurrent architectures include the neural probabilistic language model (Bengio et al., 2003) and the log-bilinear language model (Mnih and Hinton, 2007). Much more detail on these models can be found in the text by Goodfellow et al. (2016).

## 5.4 Evaluating language models

Because language models are typically components of larger systems — language modeling is not usually an application itself — we would prefer **extrinsic evaluation**. This means evaluating whether the language model improves performance on the application task, such as machine translation or speech recognition. But this is often hard to do, and depends on details of the overall system which may be irrelevant to language modeling. In contrast, **intrinsic evaluation** is task-neutral. Better performance on intrinsic metrics may be expected to improve extrinsic metrics across a variety of tasks, unless we are overoptimizing the intrinsic metric. We will discuss intrinsic metrics here, but bear in mind that it is important to also perform extrinsic evaluations to ensure that the improvements obtained on these intrinsic metrics really carry over to the applications that we care about.

### 5.4.1 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,

$$\ell(\boldsymbol{w}) = \sum_{m=1}^{M} \log \mathrm{p}(w_m \mid w_{m-1}, \ldots), \tag{5.48}$$

treating the entire held-out corpus as a single stream of tokens.

Typically, unknown words are mapped to the $\langle \text{UNK} \rangle$ token. This means that we have to estimate some probability for $\langle \text{UNK} \rangle$ on the training data. One way to do this is to fix the vocabulary $\mathcal{V}$ to the $V - 1$ words with the highest counts in the training data, and then convert all other tokens to $\langle \text{UNK} \rangle$. Other strategies for dealing with out-of-vocabulary terms are discussed in § 5.5.

### 5.4.2 Perplexity

Held-out likelihood is usually presented as **perplexity**, which is a deterministic transformation of the log-likelihood into an information-theoretic quantity,

$$\text{Perplex}(\boldsymbol{w}) = 2^{-\frac{\ell(\boldsymbol{w})}{M}}, \quad [5.49]$$

where $M$ is the total number of tokens in the held-out corpus.

Lower perplexities correspond to higher likelihoods, so lower scores are better on this metric. (How to remember: lower perplexity is better, because your language model is less perplexed.) To understand perplexity, here are some special cases:

- In the limit of a perfect language model, probability 1 is assigned to the held-out corpus, with $\text{Perplex}(\boldsymbol{w}) = 2^{-\frac{1}{M}\log_2 1} = 2^0 = 1$.
- In the opposite limit, probability zero is assigned to the held-out corpus, which corresponds to an infinite perplexity, $\text{Perplex}(\boldsymbol{w}) = 2^{-\frac{1}{M}\log_2 0} = 2^\infty = \infty$.
- Assume a uniform, unigram model in which $\text{p}(w_i) = \frac{1}{V}$ for all words in the vocabulary. Then,

$$\begin{aligned}
\log_2(\boldsymbol{w}) &= \sum_{m=1}^{M} \log_2 \frac{1}{V} = -\sum_{m=1}^{M} \log_2 V = -M \log_2 V \\
\text{Perplex}(\boldsymbol{w}) &= 2^{\frac{1}{M} M \log_2 V} \\
&= 2^{\log_2 V} \\
&= V.
\end{aligned}$$

  This is the "worst reasonable case" scenario, since you could build such a language model without even looking at the data.

In practice, $n$-gram language models tend to give perplexities in the range between 1 and $V$. For example, Jurafsky and Martin estimate a language model over a vocabularly of roughly $20,000$ words, on 38 million tokens of text from the Wall Street Journal (Jurafsky and Martin, 2009, page 97). They report the following perplexities on a held-out set of 1.5 million tokens:

- Unigram ($n = 1$): 962
- Bigram ($n = 2$): 170
- Trigram ($n = 3$): 109

Will this trend continue?

## 5.5 Out-of-vocabulary words

Through this chapter, we have assumed a **closed-vocabulary** setting — the vocabulary $\mathcal{V}$ is assumed to be a finite set. In realistic application scenarios, this assumption may not hold. Consider, for example, the problem of translating newspaper articles. The following sentence appeared in a Reuters article on January 6, 2017:[5]

> The report said U.S. intelligence agencies believe Russian military intelligence, the **GRU**, used intermediaries such as **WikiLeaks**, **DCLeaks.com** and the **Guccifer** 2.0 "persona" to release emails...

Suppose that you trained a language model on the Gigaword corpus,[6] which was released in 2003. The bolded terms either did not exist at this date, or were not widely known; they are unlikely to be in the vocabulary. The same problem can occur for a variety of other terms: new technologies, previously unknown individuals, new words (e.g., *hashtag*), and numbers.

One solution is to simply mark all such terms with a special token, $\langle \text{UNK} \rangle$. While training the language model, we decide in advance on the vocabulary (often the $K$ most common terms), and mark all other terms in the training data as $\langle \text{UNK} \rangle$. If we do not want to determine the vocabulary size in advance, an alternative approach is to simply mark the first occurrence of each word type as $\langle \text{UNK} \rangle$.

In some scenarios, we may prefer to make distinctions about the likelihood of various unknown words. This is particularly important in languages that have rich morphological systems, with many inflections for each word. For example, Spanish is only moderately complex from a morphological perspective, yet each verb has dozens of inflected forms. In such languages, there will necessarily be many word types that we do not encounter in a corpus, which are nonetheless predictable from the morphological rules of the language. To use a somewhat contrived English example, if *transfenestrate* is in the vocabulary, our language model should assign a non-zero probability to the past tense *transfenestrated*, even if it does not appear in the training data.

One way to accomplish this is to supplement word-level language models with **character-level language models**. Such models can use $n$-grams or RNNs, but with a fixed vocabulary equal to the set of ASCII or Unicode characters. For example Ling et al. (2015) propose an LSTM model over characters, and Kim (2014) employ a **convolutional neural network** (LeCun and Bengio, 1995). A more linguistically motivated approach is to segment words into meaningful subword units, known as **morphemes** (see chapter 8). For

[5] Bayoumy, Y. and Strobel, W. (2017, January 6). U.S. intel report: Putin directed cyber 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`

example, Botha and Blunsom (2014) induce vector representations for morphemes, which they build into a log-bilinear language model; Bhatia et al. (2016) incorporate morpheme vectors into an LSTM.

## Exercises

1. exercises tk

# Chapter 6

# Sequence labeling

In sequence labeling, we want to assign tags to words, or more generally, we want to assign discrete labels to elements in a sequence. There are many applications of sequence labeling in natural language processing, and chapter 7 presents an overview. One of the most classic application of sequence labeling is **part-of-speech tagging**, which involves tagging each word by its grammatical category. Coarse-grained grammatical categories include **N**OUNs, which describe things, properties, or ideas, and **V**ERBs, which describe actions and events. Given a simple sentence like,

(6.1) They can fish.

we would like to produce the tag sequence N V V, with the modal verb *can* labeled as a verb in this simplified example.

## 6.1 Sequence labeling as classification

One way to solve tagging problems is to treat them as classification. We can write $\boldsymbol{f}((\boldsymbol{w}, m), y)$ to indicate the feature function for applying tag $y$ to word $w_m$ in the sequence $w_1, w_2, \ldots, w_M$. A simple tagging model would have a single base feature, the word itself:

$$\boldsymbol{f}((\boldsymbol{w} = \textit{they can fish}, m = 1), \text{N}) = \langle \textit{they}, \text{N} \rangle \quad [6.1]$$
$$\boldsymbol{f}((\boldsymbol{w} = \textit{they can fish}, m = 2), \text{V}) = \langle \textit{can}, \text{V} \rangle \quad [6.2]$$
$$\boldsymbol{f}((\boldsymbol{w} = \textit{they can fish}, m = 3), \text{V}) = \langle \textit{fish}, \text{V} \rangle. \quad [6.3]$$

Here the feature function takes three arguments as input: the sentence to be tagged (*they can fish* in all cases), the proposed tag (e.g., N or V), and the word token to which this tag is applied. This simple feature function then returns a single feature: a tuple including the word to be tagged and the tag that has been proposed. If the vocabulary size is $V$

and the number of tags is $K$, then there are $V \times K$ features. Each of these features must be assigned a weight. These weights can be learned from a labeled dataset using a classification algorithm such as perceptron, but this isn't necessary in this case: it would be equivalent to define the classification weights directly, with $\theta_{w,y} = 1$ for the tag $y$ most 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 can go wrong. Consider the word *fish*, which often describes an animal rather than an activity; in these cases, *fish* should be tagged as a noun. To tag ambiguous words correctly, the tagger must rely on context, such as the surrounding words. We can build this context into the feature set by incorporating the surrounding words as additional features:

$$\begin{aligned}
\boldsymbol{f}((\boldsymbol{w} = \textit{they can fish}, 1), \text{N}) = \{&\langle w_i = \textit{they}, y_i = \text{N}\rangle, \\
&\langle w_{i-1} = \lozenge, y_i = \text{N}\rangle, \\
&\langle w_{i+1} = \textit{can}, y_i = \text{N}\rangle\} \qquad [6.4] \\
\boldsymbol{f}((\boldsymbol{w} = \textit{they can fish}, 2), \text{V}) = \{&\langle w_i = \textit{can}, y_i = \text{V}\rangle, \\
&\langle w_{i-1} = \textit{they}, y_i = \text{V}\rangle, \\
&\langle w_{i+1} = \textit{fish}, y_i = \text{V}\rangle\} \qquad [6.5] \\
\boldsymbol{f}((\boldsymbol{w} = \textit{they can fish}, 3), \text{V}) = \{&\langle w_i = \textit{fish}, y_i = \text{V}\rangle, \\
&\langle w_{i-1} = \textit{can}, y_i = \text{V}\rangle, \\
&\langle w_{i+1} = \blacklozenge, y_i = \text{V}\rangle\}. \qquad [6.6]
\end{aligned}$$

These features contain enough information that a tagger should be able to choose the right label for the word *fish*: words that follow the modal verb *can* are likely to be verbs themselves, so the feature $\langle w_{i-1} = \textit{can}, y_i = \text{V}\rangle$ should have a large positive weight.

However, even with this enhanced feature set, it may be difficult to tag some sequences correctly. One reason is that there are often relationships between the tags themselves. For example, in English it is relatively rare for a verb to follow another verb — particularly if we differentiate **M**ODAL verbs like *can* and *should* from more typical verbs, like *give*, *transcend*, and *befuddle*. We would like to incorporate preferences **against** such tag sequences, and preferences **for** other tag sequences, such as NOUN-VERB.

The need for such preferences is best illustrated by a **garden path sentence**:

(6.2) The old man the boat.

Grammatically, the word *the* is a **D**ETERMINER. When you read the sentence, what part of speech did you first assign to *old*? Typically, this word is an ADJECTIVE — abbreviated as J — which is a class of words that modify nouns. Similarly, *man* is usually a noun. The resulting sequence of tags is D J N D N. But this is a mistaken "garden path" interpretation, which ends up leading nowhere. It is unlikely that a determiner would directly follow a noun,[1] and particularly unlikely that the entire sentence would lack a verb. The

[1] The main exception is the double object construction, as in *I gave the child a toy.*

only possible verb in the sentence 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 can 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.

## 6.2 Sequence labeling as structure prediction

As an alternative, we can think of the entire sequence of tags as a label itself. For a given sequence of words $\boldsymbol{w}_{1:M} = (w_1, w_2, \ldots, w_M)$, there is a set of possible taggings $\mathcal{Y}(\boldsymbol{w}_{1:M}) = \mathcal{Y}^M$, where $\mathcal{Y} = \{\text{N}, \text{V}, \text{D}, \ldots\}$ 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 $\mathcal{Y}(\boldsymbol{w}_{1:M})$,

$$\hat{\boldsymbol{y}}_{1:M} = \underset{\boldsymbol{y}_{1:M} \in \mathcal{Y}(\boldsymbol{w}_{1:M})}{\text{argmax}} \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}_{1:M}, \boldsymbol{y}_{1:M}), \tag{6.7}$$

where $\boldsymbol{y}_{1:M} = (y_1, y_2, \ldots, y_M)$ is a sequence of $M$ tags. Note that in this formulation, we have a feature function that consider the entire tag sequence $\boldsymbol{y}_{1:M}$. Such a feature function can therefore 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 $w_1, \ldots, w_M$, 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 contemporary 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 $|\mathcal{Y}(\boldsymbol{w}_{1:M})| \approx 10^{20}$, 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 feature function. Suppose we choose features that never consider more than one tag. We can indicate this restriction as,

$$\boldsymbol{f}(\boldsymbol{w}, \boldsymbol{y}) = \sum_{m=1}^{M} \boldsymbol{f}(\boldsymbol{w}, y_m, m), \tag{6.8}$$

where we use the shorthand $\boldsymbol{w} \triangleq \boldsymbol{w}_{1:M}$. The summation in Equation 6.8 means that the overall feature vector is the sum of feature vectors associated with each individual tagging decision. These features are not capable of capturing the intuitions that might help us solve garden path sentences, such as the insight that determiners rarely follow

nouns in English. But this restriction does make it possible to find the globally optimal tagging, by making a sequence of individual tagging decisions.

$$\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, \boldsymbol{y}) = \boldsymbol{\theta} \cdot \sum_{m=1}^{M} \boldsymbol{f}(\boldsymbol{w}, y_m, m) \tag{6.9}$$

$$= \sum_{m=1}^{M} \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, y_m, m) \tag{6.10}$$

$$\hat{y}_m = \underset{y_m}{\text{argmax}}\, \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, y_m, m) \tag{6.11}$$

$$\hat{\boldsymbol{y}} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_M) \tag{6.12}$$

Note that we are still searching over an exponentially large set of tag sequences! But the feature set restriction results in decoupling the labeling decisions that were previous interconnected. As a result, it is not necessary to score every one of the $|\mathcal{Y}|^M$ tag sequences individually — we can find the optimal sequence by scoring the local parts of these decisions.

Now let's consider a slightly less restrictive feature function: rather than considering only individual tags, we will consider adjacent tags too. This means that we can have negative weights for infelicitous tag pairs, such as noun-determiner, and positive weights for typical tag pairs, such as determiner-noun and noun-verb. We define this feature function as,

$$\boldsymbol{f}(\boldsymbol{w}, \boldsymbol{y}) = \sum_{m=1}^{M} \boldsymbol{f}(\boldsymbol{w}, y_m, y_{m-1}, m). \tag{6.13}$$

Let's apply this feature function to the shorter example, *they can fish*, using features for word-tag and tag-tag pairs:

$$\boldsymbol{f}(\boldsymbol{w} = \textit{they can fish}, \boldsymbol{y} = \text{N V V}) = \sum_{m=1}^{M} \boldsymbol{f}(\boldsymbol{w}, y_m, y_{m-1}, m) \tag{6.14}$$

$$= \boldsymbol{f}(\boldsymbol{w}, \text{N}, \lozenge, 1) + \boldsymbol{f}(\boldsymbol{w}, \text{V}, \text{N}, 2) + \boldsymbol{f}(\boldsymbol{w}, \text{V}, \text{V}, 3) \tag{6.15}$$

$$= \langle w_m = \textit{they}, y_m = \text{N} \rangle + \langle y_m = \text{N}, y_{m-1} = \lozenge \rangle + \langle w_m = \textit{can}, y_m = \text{V} \rangle + \langle y_m = \text{V}, y_{m-1} = \text{N} \rangle + \langle w_m = \textit{fish}, y_m = \text{V} \rangle + \langle y_m = \text{V}, y_{m-1} = \text{V} \rangle + \langle y_m = \blacklozenge, y_{m-1} = \text{V} \rangle. \tag{6.16}$$

We end up with seven active features: one for each word-tag pair, and one for each tag-tag pair (this includes a final tag $y_{M+1} = \blacklozenge$). These features capture what are arguably the

two main sources of information for part-of-speech tagging: which tags are appropriate for each word, and which tags tend to follow each other in sequence. Given appropriate weights for these features, we can expect to make the right tagging decisions, even for difficult cases like *the old man the boat*.

The example shows that even with the restriction to the feature set shown in Equation 6.13, it is still possible to construct expressive features that are capable of solving many sequence labeling problems. But the key question is: does this restriction make it possible to perform efficient inference? The answer is yes, and the solution is the **Viterbi algorithm** (Viterbi, 1967).

## 6.3 The Viterbi algorithm

We now consider the inference problem,

$$\hat{\boldsymbol{y}} = \operatorname*{argmax}_{\boldsymbol{y}} \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, \boldsymbol{y}) \tag{6.17}$$

$$\boldsymbol{f}(\boldsymbol{w}, \boldsymbol{y}) = \sum_{m=1}^{M} \boldsymbol{f}(\boldsymbol{w}, y_m, y_{m-1}, m). \tag{6.18}$$

Given this restriction on the feature function, we can solve this inference problem using **dynamic programming**, a algorithmic technique for reusing work in recurrent computations. As is often the case in dynamic programming, we begin by solving an auxiliary problem: rather than finding the best tag sequence, we simply try to compute the **score** of the best tag sequence,

$$\max_{\boldsymbol{y}} \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, \boldsymbol{y}) = \max_{\boldsymbol{y}_{1:M}} \sum_{m=1}^{M} \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, y_m, y_{m-1}, m) \tag{6.19}$$

$$= \max_{\boldsymbol{y}_{1:M}} \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, y_M, y_{M-1}, M) + \sum_{m=1}^{M-1} \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, y_m, y_{m-1}, m) \tag{6.20}$$

$$= \max_{y_M} \max_{y_{M-1}} \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, y_M, y_{M-1}, M) + \max_{\boldsymbol{y}_{1:M-2}} \sum_{m=1}^{M-1} \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, y_m, y_{m-1}, m). \tag{6.21}$$

In this derivation, we first removed the final element $\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, y_M, y_{M-1}, M)$ from the sum over the sequence, and then we adjusted the scope of the the $\max$ operation, since the elements $(y_1 \dots y_{M-2})$ are irrelevant to the final term.

Let us now define the **Viterbi variable**,

$$v_m(k) \triangleq \max_{\boldsymbol{y}_{1:m-1}} \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, k, y_{m-1}, m) + \sum_{n=1}^{m-1} \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, y_n, y_{n-1}, n), \tag{6.22}$$

where lower-case $m$ indicates any position in the sequence, and $k \in \mathcal{Y}$ indicates a tag for that position. The variable $v_m(k)$ represents the score of the best tag sequence $(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_m)$ that terminates in $\hat{y}_m = k$. From this definition, we can compute the score of the best tagging of the sequence by plugging the Viterbi variables $v_M(\cdot)$ into Equation 6.21,

$$\max_{\boldsymbol{y}} \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, \boldsymbol{y}) = \max_k v_M(k). \tag{6.23}$$

Now, let us look more closely at how we can compute these Viterbi variables.

$$\begin{aligned}
v_m(k) &\triangleq \max_{\boldsymbol{y}_{1:m-1}} \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, k, y_{m-1}, m) + \sum_{n=1}^{m-1} \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, y_n, y_{n-1}, n) && [6.24] \\
&= \max_{y_{m-1}} \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, k, y_{m-1}, m) \\
&\quad + \max_{\boldsymbol{y}_{1:m-2}} \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, y_{m-1}, y_{m-2}) + \sum_{n=1}^{m-2} \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, y_n, y_{n-1}, n) && [6.25] \\
&= \max_{y_{m-1}} \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, k, y_{m-1}, m) + v_{m-1}(y_{m-1}) && [6.26] \\
v_1(y) &= \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, y, \lozenge, 1). && [6.27]
\end{aligned}$$

Equation 6.26 is a **recurrence** for computing the Viterbi variables: each $v_m(k)$ can be computed in terms of $v_{m-1}(\cdot)$, and so on. We can therefore step forward through the sequence, computing first all variables $v_1(\cdot)$ from Equation 6.27, and then computing all variables $v_2(\cdot)$, $v_3(\cdot)$, and so on, until we reach the final set of variables $v_M(\cdot)$.

Graphically, it is customary to arrange these variables in a matrix, with the sequence index $m$ on the columns, and the tag index $k$ on the rows. In this representation, each $v_{m-1}(k)$ is connected to each $v_m(k')$, forming a **trellis**, as shown in Figure 6.1. As shown in the figure, special nodes are set aside for the start and end states.

Our real goal is to find the best scoring sequence, not simply to compute its score. But as is often the case in dynamic programming, solving the auxiliary problem gets us almost all the way to our original goal. Recall that each $v_m(k)$ represents the score of the best tag sequence ending in that tag $k$ in position $m$. To compute this, we maximize over possible values of $y_{m-1}$. If we keep track of the tag that maximizes this choice at each step, then we can walk backwards from the final tag, and recover the optimal tag sequence. This is indicated in Figure 6.1 by the solid blue lines, which we trace back from the final position.

**Algorithm 10** The Viterbi algorithm.

**for** $k \in \{0, \ldots K\}$ **do**
  $v_1(k) = \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, k, \lozenge, m)$
**for** $m \in \{2, \ldots, M\}$ **do**
  **for** $k \in \{0, \ldots, K\}$ **do**
    $v_m(k) = \max_{k'} \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, k, k', m) + v_{m-1}(k')$
    $b_m(k) = \text{argmax}_{k'} \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, k, k', m) + v_{m-1}(k')$
$y_M = \text{argmax}_k \, v_M(k) + \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, \blacklozenge, k, M+1)$
**for** $m \in \{M-1, \ldots 1\}$ **do**
  $y_m = b_m(y_{m+1})$
**return** $\boldsymbol{y}_{1:M}$

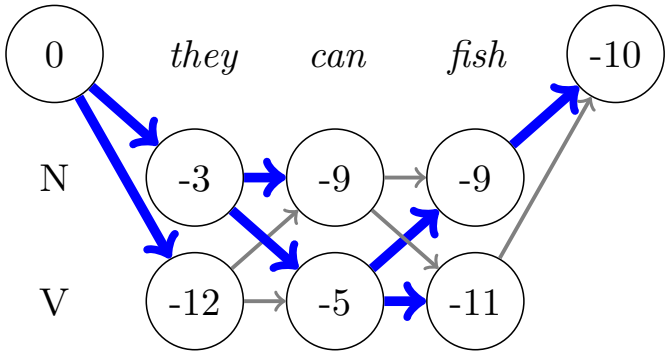



Figure 6.1: The trellis representation of the Viterbi variables, for the example *they can fish*, using the weights shown in Table 6.1.

These "back-pointers" are written $b_m(k)$, indicating the optimal tag $y_{m-1}$ on the path to $Y_m = k$.

Why does this work? We can make an inductive argument. Suppose $k$ is indeed the optimal tag for word $m$, and we now need to decide on the tag $y_{m-1}$. Because we make the inductive assumption that we know $y_m = k$, and because the feature function is restricted to adjacent tags, we need not consider any of the tags $\boldsymbol{y}_{m+1:M}$; these tags, and the features that describe them, are irrelevant to the inference of $y_{m-1}$, given that we have $y_m = k$. Thus, we are looking for the tag $\hat{y}_{m-1}$ that maximizes,

$$\hat{y}_{m-1} = \underset{y_{m-1}}{\text{argmax}} \, \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, k, y_{m-1}, m) + \max_{\boldsymbol{y}_{1:m-2}} \sum_{n=1}^{m-1} \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, y_n, y_{n-1}, n) \quad [6.28]$$

$$= \underset{y_{m-1}}{\text{argmax}} \, \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, k, y_{m-1}, m) + v_{m-1}(y_{m-1}), \quad [6.29]$$

| | *they* | *can* | *fish* |
|---|---|---|---|
| N | $-2$ | $-3$ | $-3$ |
| V | $-10$ | $-1$ | $-3$ |

(a) Weights for emission features.

| | N | V | ♦ |
|---|---|---|---|
| ◊ | $-1$ | $-2$ | $-\infty$ |
| N | $-3$ | $-1$ | $-12$ |
| 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 6.1: Feature weights for the example trellis shown in Figure 6.1. Emission weights from ◊ and ♦ are implicitly set to $-\infty$.

which we obtain by plugging in the definition of the Viterbi variable. The value $\hat{y}_{m-1}$ was identified during forward pass, when computing the value of the Viterbi variable $v_m(k)$.

The complete Viterbi algorithm is shown in Algorithm 10. This formalizes the recurrences that were described in the previous paragraphs, and handles the boundary conditions at the start and end of the sequence. Specifically, when computing the initial Viterbi variables $v_1(\cdot)$, we use a special tag, ◊, to indicate the start of the sequence. When computing the final tag $Y_M$, we use another special tag, ♦, to indicate the end of the sequence. These special tags enable the use of transition features for the tags that begin and end the sequence: for example, conjunctions are unlikely to end sentences in English, so we would like a large negative weight for the feature $\langle \text{CC}, ♦ \rangle$; nouns are relatively likely to appear at the beginning of sentences, so we would like a more positive (or less negative) weight for the feature $\langle ◊, \text{N} \rangle$.

What is the complexity of this algorithm? If there are $K$ tags and $M$ positions in the sequence, then there are $M \times K$ Viterbi variables to compute. Computing each variable requires finding a maximum over $K$ possible predecessor tags. The total computation cost of populating the trellis is therefore $\mathcal{O}(MK^2)$, with an additional factor for the number of active features at each position. After completing the trellis, we simply trace the backwards pointers to the beginning of the sequence, which takes $\mathcal{O}(M)$ operations.

### 6.3.1 Example

To illustrate the Viterbi algorithm with an example, let us consider the minimal tagset $\{\text{N}, \text{V}\}$, corresponding to nouns and verbs. Even in this tagset, there is considerable ambiguity: for example, the words *can* and *fish* can each take both tags. Of the $2 \times 2 \times 2 = 8$ possible taggings for the sentence *they can fish*, four are possible given these possible tags, and two are grammatical. (The tagging *they*/N *can*/V *fish*/N corresponds to the scenario of putting fish into cans.)

To begin, we use the feature weights defined in Table 6.1. These weights are used to

incrementally fill in the trellis. As described in Algorithm 10, we fill in the cells from left to right, with each column corresponding to a word in the sequence. As we fill in the cells, we must keep track of the back-pointers $b_m(k)$ — the previous cell that maximizes the score of tag $k$ at word $m$. These are represented in the figure with the thick blue lines. At the end of the algorithm, we recover the optimal tag sequence by tracing back the optimal path from the final position, $(M+1, \blacklozenge)$.

### 6.3.2 Higher-order features

The Viterbi algorithm was made possible by a restriction of the features to consider only pairs of adjacent tags. In a sense, we can think of this as a bigram language model, at the tag level. A natural question is how to generalize Viterbi to tag trigrams, which would involve the following feature decomposition:

$$\boldsymbol{f}(\boldsymbol{w}, \boldsymbol{y}) = \sum_{m}^{M} \boldsymbol{f}(\boldsymbol{w}, y_m, y_{m-1}, y_{m-2}, m). \quad [6.30]$$

One possibility is to take the Cartesian product of the tagset with itself, $\mathcal{Y}^{(2)} = \mathcal{Y} \times \mathcal{Y}$. The tags in this product space are ordered pairs, representing adjacent tags at the token level: for example, the tag $\langle \text{N}, \text{V} \rangle$ would represent a noun followed by a verb. Transitions between such tags must be consistent: we can have a transition from $\langle \text{N}, \text{V} \rangle$ to $\langle \text{V}, \text{N} \rangle$ (corresponding to the token-level tag sequence N V N), but not from $\langle \text{N}, \text{V} \rangle$ to $\langle \text{N}, \text{N} \rangle$, which would not correspond to any token-level tag sequence. This constraint can be enforced in the feature weights, with $\theta_{\langle\langle a,b\rangle,\langle c,d\rangle\rangle} = -\infty$ if $b \neq c$. The remaining feature weights can encode preferences for and against various tag trigrams.

In the Cartesian product tag space, there are $K^2$ tags, suggesting that the time complexity will increase to $\mathcal{O}(MK^4)$. However, it is unnecessary to $\max$ over predecessor tag bigrams that are incompatible with the current tag bigram. By exploiting these constraints, it is possible to limit the time complexity to $\mathcal{O}(MK^3)$. The space complexity is $\mathcal{O}(MK^2)$. In general, the time and space complexity of higher-order Viterbi grows exponentially with the order of the tag $n$-grams that are considered in the feature decomposition.

## 6.4 Hidden Markov Models

We now consider how to learn the weights $\boldsymbol{\theta}$ that parametrize the Viterbi sequence labeling algorithm. We begin with a probabilistic approach. Recall that the probabilistic Naïve Bayes classifier selects the label $y$ to maximize $\text{p}(y \mid \boldsymbol{x}) \propto \text{p}(y, \boldsymbol{x})$. In probabilistic sequence labeling, our goal is similar: select the tag sequence that maximizes $\text{p}(\boldsymbol{y} \mid \boldsymbol{w}) \propto \text{p}(\boldsymbol{y}, \boldsymbol{w})$. Just as Naïve Bayes could be cast as a linear classifier maximizing $\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}, y)$, we can cast

our probabilistic classifier as a linear decision rule. Furthermore, the feature restriction in Equation 6.13 can be viewed as a conditional independence assumption on the random variables $\boldsymbol{y}$. Thanks to this assumption, it is possible to perform inference using the Viterbi algorithm.

Naïve Bayes was introduced as a generative model — a probabilistic story that explains the observed data as well as the hidden label. A similar story can be constructed for probabilistic sequence labeling: first, we draw the tags from a prior distribution, $\boldsymbol{y} \sim \text{p}(\boldsymbol{y})$; next, we draw the tokens from a conditional likelihood distribution, $\boldsymbol{w} \mid \boldsymbol{y} \sim \text{p}(\boldsymbol{w} \mid \boldsymbol{y})$. However, for inference to be tractable, additional independence assumptions are required. Here we make two assumptions. First, the probability of each token depends only on its tag, and not on any other element in the sequence:

$$\text{p}(\boldsymbol{w} \mid \boldsymbol{y}) = \prod_{m=1}^{M} \text{p}(w_m \mid y_m). \quad [6.31]$$

Next, we introduce an independence assumption on the form of the prior distribution over labels: each label $y_m$ depends only on its predecessor,

$$\text{p}(\boldsymbol{y}) = \prod_{m=1}^{M} \text{p}(y_m \mid y_{m-1}), \quad [6.32]$$

where $y_0 = \lozenge$ in all cases. Due to this **Markov** assumption, probabilistic sequence labeling models are known as **hidden Markov models** (HMMs). We now state the generative model under these independence assumptions,

- For $m \in (1, 2, \ldots, M)$,
  - draw $y_m \mid y_{m-1} \sim \text{Categorical}(\boldsymbol{\lambda}_{y_{m-1}})$;
  - draw $w_m \mid y_m \sim \text{Categorical}(\boldsymbol{\phi}_{y_m})$

This generative story formalizes the hidden Markov model. Given the parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\phi}$, we can compute $\text{p}(\boldsymbol{w}, \boldsymbol{y})$ for any token sequence $\boldsymbol{w}$ and tag sequence $\boldsymbol{y}$. The HMM is often represented as a **graphical model** (Wainwright and Jordan, 2008), as shown in Figure 6.2. This representation makes the independence assumptions explicit: if a variable $v_1$ is probabilistically conditioned on another variable $v_2$, then there is an arrow $v_2 \rightarrow v_1$ in the diagram. If there are no arrows between $v_1$ and $v_2$, they are **conditionally independent**, given each variable's **Markov blanket**. In the hidden Markov model, the Markov blanket for each tag $y_m$ includes the "parent" $y_{m-1}$, and the "children" $y_{m+1}$ and $w_m$.[2]

[2] In general graphical models, a variable's Markov blanket includes its parents, children, and its children's other parents (Murphy, 2012).

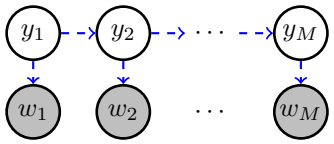



Figure 6.2: Graphical representation of the hidden Markov model. Arrows indicate probabilistic dependencies.

It is important to reflect on the implications of the HMM independence assumptions. A non-adjacent pair of tags $y_m$ and $y_n$ are conditionally independent; if $m < n$ and we are given $y_{n-1}$, then $y_m$ offers no additional information about $y_n$. However, if we are not given any information about the tags in a sequence, then all tags are probabilistically coupled.

### 6.4.1 Estimation

The hidden Markov model has two groups of parameters:

**Emission probabilities.** The probability $\mathrm{p}_e(w_m \mid y_m; \phi)$ is the emission probability, since the words are treated as probabilistically "emitted", conditioned on the tags.

**Transition probabilities.** The probability $\mathrm{p}_t(y_m \mid y_{m-1}; \lambda)$ is the transition probability, since it assigns probability to each possible tag-to-tag transition.

Both of these groups of parameters are typically computed from relative frequency estimation on a labeled corpus,

$$\phi_{k,i} \triangleq \Pr(W_m = i \mid Y_m = k) = \frac{\text{count}(W_m = i, Y_m = k)}{\text{count}(Y_m = k)}$$
$$\lambda_{k,k'} \triangleq \Pr(Y_m = k' \mid Y_{m-1} = k) = \frac{\text{count}(Y_m = k', Y_{m-1} = k)}{\text{count}(Y_{m-1} = k)}.$$

Smoothing is more important for the emission probability than the transition probability, because the event space is much larger. Smoothing techniques such as additive smoothing, interpolation, and backoff (see chapter 5) can all be applied here.

### 6.4.2 Inference

The goal of inference in the hidden Markov model is to find the highest probability tag sequence,

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\text{argmax}}\, \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{w}). \quad [6.33]$$

As in Naïve Bayes, it is equivalent to find the tag sequence with the highest **log**-probability, since the log function is monotonically increasing. It is furthermore equivalent to maximize the joint probability $\text{p}(\boldsymbol{y}, \boldsymbol{w}) = \text{p}(\boldsymbol{y} \mid \boldsymbol{w}) \times \text{p}(\boldsymbol{w}) \propto \text{p}(\boldsymbol{y} \mid \boldsymbol{w})$, which is proportional to the conditional probability. Therefore, we can reformulate the inference problem as,

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\text{argmax}} \log \text{p}(\boldsymbol{y}, \boldsymbol{w}). \tag{6.34}$$

We can now apply the HMM independence assumptions:

$$\log \text{p}(\boldsymbol{y}, \boldsymbol{w}) = \log \text{p}(\boldsymbol{y}) + \log \text{p}(\boldsymbol{w} \mid \boldsymbol{y}) \tag{6.35}$$

$$= \sum_{m=1}^{M} \log \text{p}_y(y_m \mid y_{m-1}) + \log \text{p}_{w|y}(w_m \mid y_m) \tag{6.36}$$

$$= \sum_{m=1}^{M} \log \lambda_{y_m, y_{m-1}} + \log \phi_{y_m, w_m}. \tag{6.37}$$

This log probability can be rewritten as a dot product of weights and features,

$$\log \text{p}(\boldsymbol{y}, \boldsymbol{w}) = \sum_{m=1}^{M} \log \lambda_{y_m, y_{m-1}} + \log \phi_{y_m, w_m} \tag{6.38}$$

$$= \sum_{m=1}^{M} \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, y_m, y_{m-1}, m), \tag{6.39}$$

where the feature function is defined,

$$\boldsymbol{f}(\boldsymbol{w}, y_m, y_{m-1}, m) = \{\langle y_m, y_{m-1} \rangle, \langle y_m, w_m \rangle\}, \tag{6.40}$$

and the weight vector $\theta$ encodes the log-parameters $\log \boldsymbol{\lambda}$ and $\log \boldsymbol{\phi}$.

This derivation shows that HMM inference can be viewed as an application of the Viterbi decoding algorithm, given an appropriately defined feature function and weight vector. The local product $\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, y_m, y_{m-1}, m)$ can be interpreted probabilistically,

$$\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, y_m, y_{m-1}, m) = \log \text{p}_y(y_m \mid y_{m-1}) + \log \text{p}_{w|y}(w_m \mid y_m) \tag{6.41}$$

$$= \log \text{p}(y_m, w_m \mid y_{m-1}). \tag{6.42}$$

Now recall the definition of the Viterbi variables,

$$v_m(k) = \max_{\boldsymbol{y}_{1:m-1}} \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, Y_m = k, y_{m-1}, m) + \sum_{n=1}^{m-1} \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, y_n, y_{n-1}, n) \tag{6.43}$$

$$= \max_{\boldsymbol{y}_{1:m-1}} \log \text{p}_{y_m, w_m | y_{m-1}}(k, w_m \mid y_{m-1}) + \sum_{n=1}^{m-1} \log \text{p}(y_n, w_n \mid y_{n-1}) \tag{6.44}$$

$$= \max_{\boldsymbol{y}_{1:m-1}} \log \text{p}(\boldsymbol{y}_{1:m-1}, Y_m = k, \boldsymbol{w}_{1:m}). \tag{6.45}$$

In words, the Viterbi variable $v_m(k)$ is the log probability of the best tag sequence ending in $Y_m = k$, joint with the word sequence $\boldsymbol{w}_{1:m}$. The log probability of the best complete tag sequence is therefore,

$$\max_{\boldsymbol{y}_{1:M}} \log \text{p}(\boldsymbol{y}_{1:M}, \boldsymbol{w}_{1:M}) = \max_{y_M} \log \text{p}_y(\blacklozenge \mid y_M) + v_M(y_M). \quad [6.46]$$

The Viterbi algorithm can also be implemented using probabilities, rather than log probabilities. In this case, each $v_m(k)$ is equal to,

$$\begin{aligned}
v_m(k) &= \max_{\boldsymbol{y}_{1:m-1}} \text{p}(\boldsymbol{y}_{1:m-1}, Y_m = k, \boldsymbol{w}_{1:m}) && [6.47]\\
&= \max_{y_{m-1}} \text{p}(Y_m = k, w_m \mid y_{m-1}) \times \max_{\boldsymbol{y}_{1:m-2}} \text{p}(y_{1:m-2}, y_{m-1}, \boldsymbol{w}_{1:m-1}) && [6.48]\\
&= \max_{y_{m-1}} \text{p}(Y_m = k, w_m \mid y_{m-1}) \times v_{m-1}(y_{m-1}) && [6.49]\\
&= \text{p}_E(w_m \mid Y_m = k) \times \max_{y_{m-1}} \text{p}_T(y_m \mid y_{m-1}) \times v_{m-1}(y_{m-1}) && [6.50]\\
&= \max_{y_{m-1}} \exp(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, Y_m = k, y_{m-1}, m)) \times v_{m-1}(y_{m-1}). && [6.51]
\end{aligned}$$

In the final line, we use the fact that $\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, y_m, y_{m-1}, m) = \log \text{p}(y_m, w_m \mid y_{m-1})$, and exponentiate the dot product to obtain the probability.

In practice, the probabilities tend towards zero over long sequences, so the log-probability version of Viterbi is more practical from the standpoint of numerical stability. However, this version connects to a broader literature on inference in graphical models. Each Viterbi variable is computed by **maximizing** over a set of **products**. Thus, the Viterbi algorithm is a special case of the **max-product algorithm** for inference in graphical models (Wainwright and Jordan, 2008).

### 6.4.3 The Forward Algorithm

In an influential survey, Rabiner (1989) defines three problems for hidden Markov models:

**Decoding** Find the best tags $\boldsymbol{y}$ for a sequence $\boldsymbol{w}$.

**Likelihood** Compute the marginal probability $\text{p}(\boldsymbol{w}) = \sum_{\boldsymbol{y}} \text{p}(\boldsymbol{w}, \boldsymbol{y})$.

**Learning** Given only unlabeled data $\{\boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}, \ldots, \boldsymbol{w}^{(N)}\}$, estimate the transition and emission distributions.

The Viterbi algorithm solves the decoding problem. We'll talk about the learning problem in § 6.6. Let's now consider how to compute the marginal likelihood $\text{p}(\boldsymbol{w}) = \sum_{\boldsymbol{y}} \text{p}(\boldsymbol{w}, \boldsymbol{y})$, which involves summing over all possible tag sequences. There are at least two reasons we might want to do this:

**Language modeling** Note that the probability $\text{p}(\boldsymbol{w})$ is also computed by the language models that were discussed in chapter 5. In those language models, we used only unlabeled corpora, conditioning each token $w_m$ on previous tokens. An HMM-based language model would leverage a corpus of part-of-speech annotations, and therefore might be expected to generalize better than an n-gram language model — for example, it would be more likely to assign positive probability to a nonsense grammatical sentence like *colorless green ideas sleep furiously.*

**Tag marginals** It is often important to compute marginal probabilities of individual tags, $\text{p}(y_m \mid \boldsymbol{w}_{1:M})$. This is the probability distribution over tags for token $m$, conditioned on all of the words $\boldsymbol{w}_{1:M}$. For example, we might like the know the probability that a given word is tagged as a verb, regardless of how all the other words are tagged. We will discuss how to compute this probability in § 6.5.3, but as a preview, we will use the following form,

$$\text{p}(y_m \mid \boldsymbol{w}_{1:M}) = \frac{\text{p}(y_m, \boldsymbol{w}_{1:M})}{\text{p}(\boldsymbol{w}_{1:M})}, \tag{6.52}$$

which involves the marginal likelihood in the denominator.

We can compute the marginal likelihood using a dynamic program that is nearly identical to the Viterbi algorithm. We will use probabilities for now, and show the conversion to log-probabilities later. The core of the algorithm is to compute a set of **forward variables**,

$$\alpha_m(k) \triangleq \text{p}(Y_m = k, \boldsymbol{w}_{1:m}). \tag{6.53}$$

From this definition, we can compute the marginal likelihood by summing over the final forward variables,

$$\text{p}(\boldsymbol{w}) = \sum_{k \in \mathcal{Y}} \text{p}(Y_M = k, \boldsymbol{w}_{1:M}) \tag{6.54}$$

$$= \sum_{k \in \mathcal{Y}} \alpha_M(k). \tag{6.55}$$

To capture the probability of terminating the sequence on each possible tag $Y_M$, we can pad the end of $\boldsymbol{w}$ with an extra token ■, which can only be emitted from the stop tag ♦.

The forward variables themselves can be computed recursively,

$$\begin{aligned}
\alpha_m(k) =& \mathrm{p}(Y_m = k, \boldsymbol{w}_{1:m}) && [6.56]\\
=& \mathrm{p}(w_m \mid Y_m = k) \times \Pr(Y_m = k, \boldsymbol{w}_{1:m-1}) && [6.57]\\
=& \mathrm{p}(w_m \mid Y_m = k) \times \sum_{k' \in \mathcal{Y}} \Pr(Y_m = k, Y_{m-1} = k', \boldsymbol{w}_{1:m-1}) && [6.58]\\
=& \mathrm{p}(w_m \mid Y_m = k) \times \sum_{k' \in \mathcal{Y}} \Pr(Y_m = k \mid Y_{m-1} = k') \times \Pr(Y_{m-1} = k', \boldsymbol{w}_{1:m-1}) && [6.59]\\
=& \mathrm{p}(w_m \mid Y_m = k) \times \sum_{k' \in \mathcal{Y}} \Pr(Y_m = k \mid Y_{m-1} = k') \times \alpha_{m-1}(k') && [6.60]\\
=& \sum_{k' \in \mathcal{Y}} \exp(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}_{1:M}, Y_m = k, Y_{m-1} = k', m)) \times \alpha_{m-1}(k'). && [6.61]
\end{aligned}$$

The derivation relies on the independence assumptions in the hidden Markov model: $W_m$ depends only on $Y_m$, and $Y_m$ is conditionally independent from $W_{1:m-1}$ and all tags, given $Y_{m-1}$. We complete the derivation by introducing $Y_{m-1}$ and summing over all possible values, and by then applying the chain rule to obtain the final recursive form.

Procedurally, we compute the forward variables in just the same way as we compute the Viterbi variables: we first compute all $\alpha_1(\cdot)$, then all $\alpha_2(\cdot)$, and so on. We initialize each $\alpha_0(k) = \mathrm{p}(Y_m = k \mid Y_{m-1} = \lozenge)$, to capture the transition probability from the start symbol. Comparing Equation 6.60 to Equation 6.50, the sole difference is that instead of maximizing over possible values of $Y_{m-1}$, we sum. Just as the Viterbi algorithm is a special case of the max-product algorithm for inference in graphical models, the forward algorithm is a special case of the **sum-product** algorithm for computing marginal likelihoods.

In practice, it is numerically more stable to compute the marginal log-probability. In the log domain, the forward recurrence is,

$$\begin{aligned}
\alpha_m(k) \triangleq& \log \mathrm{p}(\boldsymbol{w}_{1:m}, Y_m = k) && [6.62]\\
=& \log \sum_{k' \in \mathcal{Y}} \exp(\log \mathrm{p}(w_m \mid Y_m = k) + \log \Pr(Y_m = k \mid Y_{m-1} = k') \\
&+ \log \mathrm{p}(\boldsymbol{w}_{1:m-1}, Y_{m-1} = k')) && [6.63]\\
=& \log \sum_{k' \in \mathcal{Y}} \exp(\log \mathrm{p}(w_m \mid Y_m = k) + \log \Pr(Y_m = k \mid Y_{m-1} = k') + \alpha_{m-1}(k')) \\
& && [6.64]\\
=& \log \sum_{k' \in \mathcal{Y}} \exp(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}_{1:M}, Y_m = k, Y_{m-1} = k', m) + \alpha_{m-1}(k')). && [6.65]
\end{aligned}$$

Scientific programming libraries provide numerically robust implementations of the log-sum-exp function, which should prevent overflow and underflow from exponentiation.

### 6.4.4 Semiring Notation and the Generalized Viterbi Algorithm

We have now seen the Viterbi and Forward recurrences, each of which can be performed over probabilities or log probabilities. These four recurrences are closely related, and can in fact be expressed as a single recurrence in a more general notation, known as **semiring algebra**. We use the symbol $\oplus$ to represent generalized addition, and the symbol $\otimes$ to represent generalized multiplication.[3] Given these operators, we can denote a generalized Viterbi recurrence as,

$$v_m(k) = \bigoplus_{k' \in \mathcal{Y}} \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, Y_m = k, Y_{m-1} = k', m) \otimes v_{m-1}(k'). \quad [6.66]$$

Each recurrence that we have seen so far is a special case of this generalized Viterbi recurrence:

- In the max-product Viterbi recurrence over probabilities, the $\oplus$ operation corresponds to maximization, and the $\otimes$ operation corresponds to multiplication.
- In the forward recurrence over probabilities, the $\oplus$ operation corresponds to addition, and the $\otimes$ operation corresponds to multiplication.
- In the max-product Viterbi recurrence over log-probabilities, the $\oplus$ operation corresponds to maximization, and the $\otimes$ operation corresponds to addition. (This is sometimes called the **tropical semiring**, in honor of the Brazilian mathematician Imre Simon.)
- In the forward recurrence over log-probabilities, the $\oplus$ operation corresponds to log-addition, $a \oplus b = \log(e^a + e^b)$. The $\otimes$ operation corresponds to addition.

The mathematical abstraction offered by semiring notation can be applied to the software implementations of these algorithms, yielding concise and modular implementations. The OpenFST library (Allauzen et al., 2007) is an example of a software package in which the algorithms are parametrized by the choice of semiring.

## 6.5 Discriminative sequence labeling

Today, hidden Markov models are rarely used for supervised sequence labeling. This is because HMMs are limited to only two phenomena:

- Word-tag probabilities, via the emission probability $\mathrm{p}_E(w_m \mid y_n)$;

[3] In 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 $\overline{0}$ and $\overline{1}$, such that $a \oplus \overline{0} = a$ and $a \otimes \overline{1} = a$; and there must be a multiplicative annihilator $\overline{0}$, such that $a \otimes \overline{0} = \overline{0}$.

- local context, via the transition probability $\mathrm{p}_T(y_m \mid y_{m-1})$.

However, as we have seen, the Viterbi algorithm can be applied to much more general feature sets, as long as the decomposition $\boldsymbol{f}(\boldsymbol{w}, \boldsymbol{y}) = \sum_{m=1}^{M} \boldsymbol{f}(\boldsymbol{w}, y_m, y_{m-1}, m)$ is observed. In this section, we discuss methods for learning the weights on such features. However, let's first pause to ask what additional features might be needed.

**Word affix features.** Consider the problem of part-of-speech tagging on the first four lines of the poem *Jabberwocky* (Carroll, 1917):

(6.3) 'Twas brillig, and the slithy toves
Did gyre and gimble in the wabe:
All mimsy were the borogoves,
And the mome raths outgrabe.

Many of these words are made up, so you would have no information about their probabilities of being associated with any particular part of speech. Yet it is not so hard to see what their grammatical roles might be in this passage. Context helps: for example, the word *slithy* follows the determiner *the*, and therefore is likely to be a noun or adjective. Which do you think is more likely? The suffix *-thy* is found in a number of adjectives — e.g., *frothy,healthy,pithy,worthy*. The suffix is also found in a handful of nouns — e.g., *apathy,sympathy* — but nearly all of these nouns contain *-pathy*, unlike *slithy*. The suffix gives some evidence that *slithy* is an adjective, and indeed it is: later in the text we find that it is a combination of the adjectives *lithe* and *slimy*.[4]

**Fine-grained context.** Another useful source of information is fine-grained context — that is, contextual information that is more specific than the previous tag. For example, consider the noun phrases *this fish* and *these fish*. Many part-of-speech tagsets distinguish between singular and plural nouns, but do not distinguish between singular and plural determiners; for example, the Penn Treebank tagset follows these conventions. A hidden Markov model would be unable to correctly label *fish* as singular or plural in both of these cases, because it only has access to two features: the preceding tag (determiner in both cases) and the word (*fish* in both cases). The classification-based tagger discussed in § 6.1 had the ability to use preceding and succeeding words as features, and we would like to incorporate this information into a sequence labeling algorithm.

[4] **morphology** is the study of how words are formed from smaller linguistic units. Computational approaches to morphological analysis are touched on in chapter 8; Bender (2013) provides a good overview of the underlying linguistic principles.

**Example** Suppose we have the tagging D J N (determiner, adjective, noun) for the sequence *the slithy toves* in Jabberwocky, so that

$$\begin{aligned}\boldsymbol{w} &= \textit{the slithy toves} \\ \boldsymbol{y} &= \text{D J N}.\end{aligned}$$

We now create the feature vector for this example, assuming that we have word-tag features (indicated by prefix $W$), tag-tag features (indicated by prefix $T$), and suffix features (indicated by prefix $M$). We assume access to a method for extracting the suffix *-thy* from *slithy*, *-es* from *toves*, and $\varnothing$ from *the*, indicating that this word has no suffix. The resulting feature vector is,

$$\begin{aligned}\boldsymbol{f}(\text{the slithy toves, D J N}) = \{&\langle W : \text{the, D}\rangle, \langle M : \varnothing, \text{D}\rangle, \langle T : \lozenge, \text{D}\rangle \\ &\langle W : \text{slithy, J}\rangle, \langle M : \text{-thy, J}\rangle, \langle T : \text{D, J}\rangle \\ &\langle W : \text{toves, N}\rangle, \langle M : \text{-es, N}\rangle, \langle T : \text{J, N}\rangle \\ &\langle T : \text{N}, \blacklozenge\rangle\}.\end{aligned}$$

We now consider several discriminative methods for learning feature weights in sequence labeling. In § 1.2, we considered three types of discriminative classifiers: perceptron, support vector machine, and logistic regression. Each of these classifiers has a structured equivalent, enabling it to be trained from labeled sequences rather than individual tokens.

### 6.5.1 Structured perceptron

The perceptron classifier updates its weights 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:

$$\hat{y} = \underset{y \in \mathcal{Y}}{\text{argmax}}\, \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}, y) \quad [6.67]$$

$$\boldsymbol{\theta}^{(t+1)} \leftarrow \boldsymbol{\theta}^{(t)} + \boldsymbol{f}(\boldsymbol{x}, y) - \boldsymbol{f}(\boldsymbol{x}, \hat{y}). \quad [6.68]$$

We can apply exactly the same update in the case of structure prediction,

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{w})}{\text{argmax}}\, \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, \boldsymbol{y}) \quad [6.69]$$

$$\boldsymbol{\theta}^{(t+1)} \leftarrow \boldsymbol{\theta}^{(t)} + \boldsymbol{f}(\boldsymbol{w}, \boldsymbol{y}) - \boldsymbol{f}(\boldsymbol{w}, \hat{\boldsymbol{y}}). \quad [6.70]$$

This learning algorithm is called **structured perceptron**, because it learns to predict the structured output $\boldsymbol{y}$. The key difference is that instead of computing $\hat{\boldsymbol{y}}$ by enumerating

the entire set $\mathcal{Y}$, we use the Viterbi algorithm to search this set efficiently. In this case, the output structure is the sequence of tags $(y_1, y_2, \ldots, y_M)$; the algorithm can be applied to other structured outputs as long as efficient inference is possible. As in perceptron classification, weight averaging is crucial to get good performance (see § 1.2.2).

**Example** For the example *They can fish*, suppose the reference tag sequence is N V V, but our tagger incorrectly returns the tag sequence N V N. Given **feature templates** $\langle w_m, y_m \rangle$ and $\langle y_{m-1}, y_m \rangle$, the corresponding structured perceptron update is:

$$\theta_{\langle \textit{fish}, \text{V} \rangle} \leftarrow \theta_{\langle \textit{fish}, \text{V} \rangle} + 1 \quad [6.71]$$
$$\theta_{\langle \textit{fish}, \text{N} \rangle} \leftarrow \theta_{\langle \textit{fish}, \text{N} \rangle} - 1 \quad [6.72]$$
$$\theta_{\langle \text{V}, \text{V} \rangle} \leftarrow \theta_{\langle \text{V}, \text{V} \rangle} + 1 \quad [6.73]$$
$$\theta_{\langle \text{V}, \text{N} \rangle} \leftarrow \theta_{\langle \text{V}, \text{N} \rangle} - 1 \quad [6.74]$$
$$\theta_{\langle \text{V}, \blacklozenge \rangle} \leftarrow \theta_{\langle \text{V}, \blacklozenge \rangle} + 1 \quad [6.75]$$
$$\theta_{\langle \text{N}, \blacklozenge \rangle} \leftarrow \theta_{\langle \text{N}, \blacklozenge \rangle} - 1. \quad [6.76]$$

### 6.5.2 Structured Support Vector Machines

Large-margin classifiers such as the support vector machine improve on the perceptron by learning weights that push the classification boundary away from the training instances. In many cases, large-margin classifiers outperform the perceptron, so we would like to apply similar ideas to sequence labeling. A support vector machine in which the output is a structured object, such as a sequence, is called a **structured support vector machine** (Tsochantaridis et al., 2004).[5]

In classification, we formalized the large-margin constraint as,

$$\forall y \neq y^{(i)}, \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}, y^{(i)}) - \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}, y) \geq 1, \quad [6.77]$$

which says that we require a margin of at least 1 between the scores for all labels $y$ that are not equal to the correct label $y^{(i)}$. The weights $\boldsymbol{\theta}$ are then learned by constrained optimization (see § 1.3.2).

We can apply this idea to sequence labeling by formulating an equivalent set of constraints for all possible labelings $\mathcal{Y}(\boldsymbol{w})$ for an input $\boldsymbol{w}$. However, there are two problems with this idea. First, in sequence labeling, some predictions are more wrong than others: we may miss only one tag out of fifty, or we may get all fifty wrong. We would like our learning algorithm to be sensitive to this difference. Second, the number of contraints is equal to the number of possible labelings, which is exponentially large in the length of the sequence.

[5] This 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.

The first problem can be addressed by adjusting the constraint to require larger margins for more serious errors. Let $c(\boldsymbol{y}^{(i)}, \hat{\boldsymbol{y}}) \geq 0$ represent the **cost** of predicting label $\hat{\boldsymbol{y}}$ when the true label is $\boldsymbol{y}^{(i)}$. We can then generalize the margin constraint,

$$\forall \boldsymbol{y} \neq \boldsymbol{y}^{(i)}, \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}^{(i)}, \boldsymbol{y}^{(i)}) - \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}^{(i)}, \boldsymbol{y}) \geq c(\boldsymbol{y}^{(i)}, \boldsymbol{y}). \tag{6.78}$$

This cost-augmented margin constraint specializes to the constraint in Equation 6.77 if we choose the delta function $c(\boldsymbol{y}^{(i)}, \boldsymbol{y}) = \delta(\boldsymbol{y}^{(i)} \neq \boldsymbol{y})$. For sequence labeling, we can instead use a structured cost function, such as the **Hamming cost**,

$$c(\boldsymbol{y}^{(i)}, \boldsymbol{y}) = \sum_{m=1}^{M} \delta(y_m^{(i)} \neq y_m). \tag{6.79}$$

With this cost function, we require that the true labeling be seperated from the alternatives by a margin that is proportional to the number of incorrect tags in each alternative labeling. Other cost functions are possible as well.

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 $\hat{\boldsymbol{y}}$ that *maximally* violates the margin constraint. We find this prediction by solving the following **cost-augmented decoding** problem:

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y} \neq \boldsymbol{y}^{(i)}}{\text{argmax}}\, \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}^{(i)}, \boldsymbol{y}) - \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}^{(i)}, \boldsymbol{y}^{(i)}) + c(\boldsymbol{y}^{(i)}, \boldsymbol{y}) \tag{6.80}$$

$$= \underset{\boldsymbol{y} \neq \boldsymbol{y}^{(i)}}{\text{argmax}}\, \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}^{(i)}, \boldsymbol{y}) + c(\boldsymbol{y}^{(i)}, \boldsymbol{y}), \tag{6.81}$$

where in the second line we drop the term $\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}^{(i)}, \boldsymbol{y}^{(i)})$, which is constant in $\boldsymbol{y}$.

We can now formulate the margin constraint for sequence labeling,

$$\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}^{(i)}, \boldsymbol{y}^{(i)}) - \max_{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{w})} \Big( \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}^{(i)}, \boldsymbol{y}) + c(\boldsymbol{y}^{(i)}, \boldsymbol{y}) \Big) \geq 0. \tag{6.82}$$

If the score for $\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}^{(i)}, \boldsymbol{y}^{(i)})$ is greater than the cost-augmented score for all alternatives, then the constraint will be met. Therefore we can maximize over the entire set $\mathcal{Y}(\boldsymbol{w})$, meaning that we can apply Viterbi directly.[6]

The name "cost-augmented decoding" is due to the fact that the objective includes the standard decoding problem, $\max_{\hat{\boldsymbol{y}}} \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, \hat{\boldsymbol{y}})$, plus an additional term for the cost. Essentially, we want to train against predictions that are strong and wrong: they should

[6] To maximize over the set $\mathcal{Y}(\boldsymbol{w}) \setminus \boldsymbol{y}^{(i)}$ we would need an alternative version of Viterbi that returns the $k$-best predictions. $K$-best Viterbi may be useful for other reasons — for example, in interactive applications, it can be helpful to show the user multiple possible taggings. The design of $k$-best Viterbi is left an exercise.

score highly according to the model, yet incur a large loss with respect to the ground truth. We can then adjust the weights to reduce the score of these predictions.

For cost-augmented decoding to be tractable, the cost function must decompose into local parts, just as the feature function $\boldsymbol{f}(\cdot)$ does. The Hamming cost, defined above, obeys this property. To solve this cost-augmented decoding problem using the Hamming cost, we can simply add features $f_m(y_m) = \delta(y_m \neq y_m^{(i)})$, and assign a weight of 1 to these features. Decoding can then be performed using the Viterbi algorithm. Are there cost functions that do not decompose into local parts? Suppose we want to assign a constant loss $c$ to any prediction $\hat{\boldsymbol{y}}$ 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.

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:

$$\min_{\boldsymbol{\theta}} \quad \frac{1}{2}||\boldsymbol{\theta}||_2^2 - C\left(\sum_i \boldsymbol{\theta}\cdot\boldsymbol{f}(\boldsymbol{w}^{(i)},\boldsymbol{y}^{(i)}) - \max_{\boldsymbol{y}\in\mathcal{Y}(\boldsymbol{w}^{(i)})}\left[\boldsymbol{\theta}\cdot\boldsymbol{f}(\boldsymbol{w}^{(i)},\boldsymbol{y}) + c(\boldsymbol{y}^{(i)},\boldsymbol{y})\right]\right), \quad [6.83]$$

In this formulation, $C$ is a parameter that controls the tradeoff between the regularization term and the margin constraints. A number of optimization algorithms have been proposed for structured support vector machines, some of which are discussed in § 1.3.2. An empirical comparison by Kummerfeld et al. (2015) shows that stochastic subgradient descent — which is relatively easy to implement — is highly competitive, especially on the sequence labeling task of named entity recognition.

### 6.5.3 Conditional random fields

Structured perceptron is easy to implement, and structured support vector machines give excellent performance. However, sometimes we need to compute probabilities over labelings, $\text{p}(\boldsymbol{y} \mid \boldsymbol{w})$, and we would like to do this in a discriminative way. The **conditional random field** (CRF; Lafferty et al., 2001) is a conditional probabilistic model for sequence labeling; just as structured perceptron is built on the perceptron classifier, conditional random fields are built on the logistic regression classifier.[7] The basic probability model is,

$$\text{p}(\boldsymbol{y} \mid \boldsymbol{w}) = \frac{\exp(\boldsymbol{\theta}\cdot\boldsymbol{f}(\boldsymbol{w},\boldsymbol{y}))}{\sum_{\boldsymbol{y}'\in\mathcal{Y}(\boldsymbol{w})}\exp(\boldsymbol{\theta}\cdot\boldsymbol{f}(\boldsymbol{w},\boldsymbol{y}'))}. \quad [6.84]$$

[7] The 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 more generally, cliques) of variables in a **factor graph**. In sequence labeling, the pairs of variables include all adjacent tags $\langle y_m, y_{m-1}\rangle$. The probability is **conditioned** on the words $\boldsymbol{w}_{1:M}$, which are always observed, motivating the term "conditional" in the name.

This is almost identical to logistic regression, but because the label space is now tag sequences, we require efficient algorithms for both **decoding** (searching for the best tag sequence given a sequence of words $\boldsymbol{w}$ and a model $\boldsymbol{\theta}$) and for **normalizing** (summing over all tag sequences). These algorithms will be based on the usual locality assumption on the feature function, $\boldsymbol{f}(\boldsymbol{w}, \boldsymbol{y}) = \sum_{m=1}^{M} \boldsymbol{f}(\boldsymbol{w}, y_m, y_{m-1}, m)$.

### Decoding in CRFs

Decoding — finding the tag sequence $\hat{\boldsymbol{y}}$ that maximizes $\text{p}(\boldsymbol{y} \mid \boldsymbol{w})$ — is a direct application of the Viterbi algorithm. The key observation is that the decoding problem does not depend on the denominator of $\text{p}(\boldsymbol{y} \mid \boldsymbol{w})$,

$$\begin{aligned}
\hat{\boldsymbol{y}} &= \operatorname*{argmax}_{\boldsymbol{y}} \text{p}(\boldsymbol{y} \mid \boldsymbol{w}) \\
&= \operatorname*{argmax}_{\boldsymbol{y}} \log \text{p}(\boldsymbol{y} \mid \boldsymbol{w}) \\
&= \operatorname*{argmax}_{\boldsymbol{y}} \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{y}, \boldsymbol{w}) - \log \sum_{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{w})} e^{\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{y}', \boldsymbol{w})} \\
&= \operatorname*{argmax}_{\boldsymbol{y}} \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{y}, \boldsymbol{w}).
\end{aligned}$$

This is identical to the decoding problem for structured perceptron, so the same Viterbi recurrence as defined in Equation 6.26 can be used.

### Learning in CRFs

As with logistic regression, we learn the weights $\boldsymbol{\theta}$ by minimizing the regularized negative log conditional probability,

$$\ell = \frac{\lambda}{2}||\boldsymbol{\theta}||^2 - \sum_{i=1}^{N} \log \text{p}(\boldsymbol{y}^{(i)} \mid \boldsymbol{w}^{(i)}; \boldsymbol{\theta}) \quad [6.85]$$

$$= \frac{\lambda}{2}||\boldsymbol{\theta}||^2 - \sum_{i=1}^{N} \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}^{(i)}, \boldsymbol{y}^{(i)}) + \log \sum_{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{w}^{(i)})} \exp\left(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}^{(i)}, \boldsymbol{y}')\right), \quad [6.86]$$

where $\lambda$ controls the amount of regularization. We will optimize $\boldsymbol{\theta}$ by moving along the gradient of this loss. Probabilistic programming environments, such as THEANO (Bergstra et al., 2010) and TORCH (Collobert et al., 2011), can compute this gradient using automatic differentiation. However, it is worth deriving the gradient to understand how this model works, and why learning is computationally tractable.

As in logistic regression, the gradient includes a difference between observed and expected feature counts:

$$\frac{d\ell}{d\theta_j} = \lambda\theta_j + \sum_{i=1}^{N} E[f_j(\boldsymbol{w}^{(i)}, \boldsymbol{y})] - f_j(\boldsymbol{w}^{(i)}, \boldsymbol{y}^{(i)}), \quad [6.87]$$

where $f_j(\boldsymbol{w}^{(i)}, \boldsymbol{y}^{(i)})$ refers to the count of feature $j$ for token sequence $\boldsymbol{w}^{(i)}$ and tag sequence $\boldsymbol{y}^{(i)}$.

The expected feature counts are computed by summing over all possible labelings of the word sequence,

$$E[f_j(\boldsymbol{w}^{(i)}, \boldsymbol{y})] = \sum_{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{w}^{(i)})} \text{p}(\boldsymbol{y} \mid \boldsymbol{w}^{(i)}; \boldsymbol{\theta}) f_j(\boldsymbol{w}^{(i)}, \boldsymbol{y}) \quad [6.88]$$

This looks bad: it is a sum over an exponential number of labelings. To solve this problem, we again rely on the assumption that the overall feature vector decomposes into a sum of local feature vectors, which we exploit to compute the expected feature counts as a sum across the sequence:

$$\begin{aligned}
E[f_j(\boldsymbol{w}, \boldsymbol{y})] &= \sum_{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{w})} \text{p}(\boldsymbol{y} \mid \boldsymbol{w}; \boldsymbol{\theta}) f_j(\boldsymbol{w}, \boldsymbol{y}) & [6.89] \\
&= \sum_{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{w})} \text{p}(\boldsymbol{y} \mid \boldsymbol{w}; \boldsymbol{\theta}) \sum_{m=1}^{M} f_j(\boldsymbol{w}, y_m, y_{m-1}, m) & [6.90] \\
&= \sum_{m=1}^{M} \sum_{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{w})} \text{p}(\boldsymbol{y} \mid \boldsymbol{w}; \boldsymbol{\theta}) f_j(\boldsymbol{w}, y_m, y_{m-1}, m) & [6.91] \\
&= \sum_{m=1}^{M} \sum_{k,k'}^{\mathcal{Y}} \sum_{\boldsymbol{y}: y_{m-1}=k', y_m=k} \text{p}(\boldsymbol{y} \mid \boldsymbol{w}; \boldsymbol{\theta}) f_j(\boldsymbol{w}, k, k', m) & [6.92] \\
&= \sum_{m=1}^{M} \sum_{k,k'}^{\mathcal{Y}} f_j(\boldsymbol{w}, k, k', m) \sum_{\boldsymbol{y}: y_{m-1}=k', y_m=k} \text{p}(\boldsymbol{y} \mid \boldsymbol{w}; \boldsymbol{\theta}) & [6.93] \\
&= \sum_{m=1}^{M} \sum_{k,k'}^{\mathcal{Y}} f_j(\boldsymbol{w}, k', k, m) \Pr(Y_{m-1} = k', Y_m = k \mid \boldsymbol{w}; \boldsymbol{\theta}). & [6.94]
\end{aligned}$$

This derivation works by interchanging the sum over tag sequences with the sum over indices $m$. At each position in the sequence, the locality restriction on features ensures that we need only the marginal probability of the tag bigram, $\Pr(Y_{m-1} = k', Y_m = k \mid \boldsymbol{w}; \boldsymbol{\theta})$. These tag bigram marginals are also used in unsupervised approaches to sequence labeling. In principle, these marginals still require a sum over the exponentially many label

sequences in which $Y_{m-1} = k'$ and $Y_m = k$. However, the marginals can be computed efficiently using the **forward-backward algorithm**.

***Forward-backward algorithm**

Recall that in the hidden Markov model, it was possible to use the forward algorithm to compute marginal probabilities $\text{p}(y_m, \boldsymbol{w}_{1:m})$. We now derive a more general version of the forward algorithm, in which label sequences are scored in terms of **potentials** $\psi_m(k, k')$:

$$\psi_m(k, k') \triangleq \exp(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, k, k', m)) \quad [6.95]$$

$$\text{p}(\boldsymbol{y} \mid \boldsymbol{w}) = \frac{\prod_{m=1}^{M} \psi_m(y_m, y_{m-1})}{\sum_{\boldsymbol{y}'} \prod_{m=1}^{M} \psi_m(y'_m, y'_{m-1})}. \quad [6.96]$$

Equation 6.96 simply expresses the CRF conditional likelihood, under a change of notation.

The tag bigram marginal probabilities can be written as,

$$\Pr(Y_{m-1} = k', Y_m = k \mid \boldsymbol{w}; \boldsymbol{\theta}) = \frac{\sum_{\boldsymbol{y}: y_m = k, y_{m-1} = k'} \prod_{n=1}^{M} \psi_n(y_n, y_{n-1})}{\sum_{\boldsymbol{y}'} \prod_{n=1}^{M} \psi_n(y'_n, y'_{n-1})}. \quad [6.97]$$

where the denominator is the marginal $\text{p}(\boldsymbol{w}_{1:M}) = \sum_{\boldsymbol{y}} \text{p}(\boldsymbol{w}, \boldsymbol{y}_{1:M})$, sometimes known as the **partition function**.[8] Let us now consider how to compute each of these terms efficiently.

**Computing the numerator** In Equation 6.97, we sum 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 this arc, we can decompose this sum into three parts: the sum over **prefixes** $\boldsymbol{y}_{1:m-1}$, the transition $(Y_{m-1} = k') \rightarrow (Y_m = k)$, and the sum over **suffixes** $\boldsymbol{y}_{m:M}$,

$$\begin{aligned} \sum_{\boldsymbol{y}: Y_m = k, Y_{m-1} = k'} \prod_{n=1}^{M} \psi_n(y_n, y_{n-1}) = & \sum_{\boldsymbol{y}_{1:m-1}: y_{m-1} = k'} \prod_{n=1}^{m-1} \psi_n(y_n, y_n - 1) \\ & \times \psi_m(k, k') \\ & \times \sum_{\boldsymbol{y}_{m:M}: y_m = k} \prod_{n=m+1}^{M} \psi_n(y_n, y_{n-1}). \end{aligned} \quad [6.98]$$

The result is product of three terms: a score for getting to the position $(Y_{m-1} = k')$, a score for the transition from $k'$ to $k$, and a score for finishing the sequence from $(Y_m = k)$.

[8] The terminology of "potentials" and "partition functions" comes from statistical mechanics (Bishop, 2006).

Let us define the first term as a **forward variable**,

$$\alpha_m(k) \triangleq \sum_{\boldsymbol{y}_{1:m}:y_m=k} \prod_{n=1}^{m} \psi_n(y_n, y_{n-1}) \tag{6.99}$$

$$= \sum_{k' \in \mathcal{Y}} \psi_m(k, k') \sum_{\boldsymbol{y}_{1:m-1}:y_{m-1}=k'} \prod_{n=1}^{m-1} \psi_n(y_n, y_{n-1}) \tag{6.100}$$

$$= \sum_{k' \in \mathcal{Y}} \psi_m(k, k') \times \alpha_{m-1}(k'). \tag{6.101}$$

Thus, we compute the forward variables while moving from left to right over the trellis. This forward recurrence is a generalization of the forward recurrence defined in § 6.4. If $\psi_m(k, k') = \text{p}_E(w_m \mid Y_m = k) \times \text{p}_T(k \mid k')$, then we exactly recover the Hidden Markov Model forward variable $\alpha_m(k) = \text{p}(\boldsymbol{w}_{1:m}, Y_m = k)$ as computed in § 6.4.3.

The third term of Equation 6.98 can also be defined recursively, this time moving over the trellis from right to left. The resulting recurrence is called the **backward algorithm**:

$$\beta_{m-1}(k) \triangleq \sum_{\boldsymbol{y}_{m-1:M}:y_{m-1}=k} \prod_{n=m}^{M} \psi_n(y_n, y_{n-1}) \tag{6.102}$$

$$= \sum_{k' \in \mathcal{Y}} \psi_m(k', k) \sum_{\boldsymbol{y}_{m:M}:y_m=k'} \prod_{n=m+1}^{M} \psi_n(y_n, y_{n-1}) \tag{6.103}$$

$$= \sum_{k' \in \mathcal{Y}} \psi_m(k', k) \times \beta_m(k'). \tag{6.104}$$

In practice, numerical stability requires that we use log-potentials rather than potentials, $\log \psi_m(y_m, y_{m-1}) = \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, y_m, y_{m-1}, m)$. Then the sums must be replaced with log-sum-exp:

$$\log \alpha_m(k) = \log \sum_{k' \in \mathcal{Y}} \exp\left(\log \psi_m(k, k') + \log \alpha_{m-1}(k')\right) \tag{6.105}$$

$$\log \beta_{m-1}(k) = \log \sum_{k' \in \mathcal{Y}} \exp\left(\log \psi_m(k', k) + \log \beta_m(k')\right). \tag{6.106}$$

Both the forward and backward algorithm operate on the trellis, which implies a space complexity $\mathcal{O}(MK)$. Because they require computing a sum over $K$ terms at each node in the trellis, their time complexity is $\mathcal{O}(MK^2)$.

**Computing the normalization term** The normalization term (partition function), sometimes abbreviated as $Z$, can be written as,

$$Z \triangleq \sum_{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{w})} \text{p}(\boldsymbol{w}, \boldsymbol{y}) \tag{6.107}$$

$$= \sum_{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{w})} \exp\left(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, \boldsymbol{y})\right) \tag{6.108}$$

$$= \sum_{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{w})} \prod_{m=1}^{M} \exp\left(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, y_m, y_{m-1}, m)\right) \tag{6.109}$$

$$= \sum_{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{w})} \prod_{m=1}^{M} \psi_m(y_m, y_{m-1}). \tag{6.110}$$

This term can be computed directly from either the forward or backward probabilities:

$$Z = \sum_{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{w})} \prod_{m=1}^{M} \psi_m(y_m, y_{m-1}) \tag{6.111}$$

$$= \alpha_{M+1}(\blacklozenge) \tag{6.112}$$

$$= \beta_0(\lozenge). \tag{6.113}$$

**CRF learning: wrapup** Having computed the forward and backward variables, we can compute the desired marginal probability as,

$$P(Y_{m-1} = k', Y_m = k \mid \boldsymbol{w}_{1:M}) = \frac{\alpha_{m-1}(k')\psi_m(k, k')\beta_m(k)}{Z}. \tag{6.114}$$

This computation is known as the **forward-backward algorithm.** From the resulting marginals, we can compute the feature expectations $E[f_j(\boldsymbol{w}, \boldsymbol{y})]$; from these expectations, we compute a gradient on the weights $\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}}$. Stochastic gradient descent or quasi-Newton optimization can then be applied. As the optimization algorithm changes the weights, the potentials change, and therefore so do the marginals. Each iteration of the optimization algorithm therefore requires recomputing the forward and backward variables for each training instance. [9]

### 6.5.4 Neural sequence labeling

Recently, neural network methods have been applied to sequence labeling. These methods can be employed in tandem with structure prediction algorithms such as Viterbi,

[9] The `CRFsuite` package implements several learning algorithms for CRFs (`http://www.chokkan.org/software/crfsuite/`).

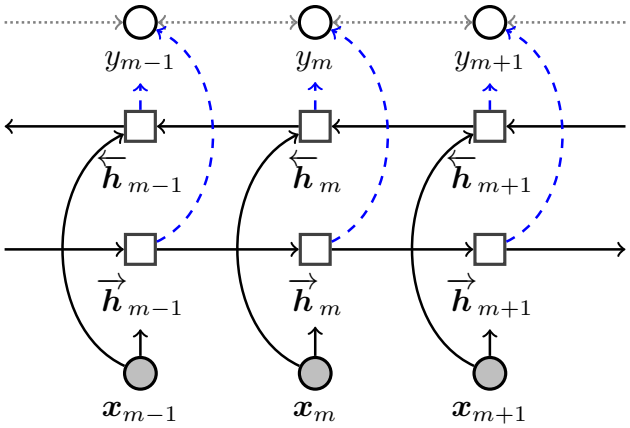



Figure 6.3: Bidirectional LSTM for sequence labeling. The solid lines indicate computation, the dashed lines indicate probabilistic dependency, and the dotted lines indicate the optional additional probabilistic dependencies between labels in the biLSTM-CRF.

although they are effective on their own. A relatively straightforward approach is to train a recurrent neural network or LSTM, as described in chapter 5; however, rather than predicting the next word in sequence, the model can be trained to predict the tag of the current word. A particularly effective approach is to train a **bidirectional recurrent neural network** (Graves and Schmidhuber, 2005), which estimates two hidden vectors, $\boldsymbol{h}_m = (\overrightarrow{\boldsymbol{h}}_m, \overleftarrow{\boldsymbol{h}}_m)$, with $\overrightarrow{\boldsymbol{h}}_m$ indicating the hidden state in a standard left-to-right RNN or LSTM, and $\overleftarrow{\boldsymbol{h}}_m$ indicating the hidden state in a left-to-right model. In this way, information from the entire sentence is brought to bear on the tagging decision for $y_m$. This approach was employed by Ling et al. (2015), who find that bi-LSTMs perform considerably better than bi-RNNs on part-of-speech tagging, and that bidirectional variants of both models perform slightly but consistently better than their unidirectional counterparts. Note that LSTM-based tagging is a classification approach, so dynamic programming is not required to find the best tag sequence (this corresponds to Figure 6.3, excluding the finely dotted lines between adjacent tags).

It is also possible in neural sequence labeling to couple the tagging decisions, by introducing additional parameters for tag-to-tag transitions. This model is called the **LSTM-CRF**, due to its combination of aspects of the long short-term memory and conditional random field models (Huang et al., 2015). The model is shown in Figure 6.3. Lample et al. (2016) find that the LSTM-CRF is especially effective on the task of **named entity recognition**, a sequence labeling task that is described in detail in § 7.3. This task has particularly strong dependencies between adjacent tags, so it is not surprising to see an advantage for structure predictions here.

Both Ling et al. (2015) and Lample et al. (2016) find it advantageous to model unseen and rare words through character-level representations. They train nested bi-LSTMs for each word, and take the concatenation of the introductory and final hidden states, $(\overrightarrow{\boldsymbol{h}}_M, \overleftarrow{\boldsymbol{h}}_0)$ as the word embeddings, which is then used as input in the tagging bi-LSTM. Lample et al. (2016) combine these character-level embeddings with **pre-trained** word embeddings, which were estimated from an unlabeled dataset that is many orders of magnitude larger than the labeled data. They do not backpropagate into the word embeddings, and learn only the transition and output parameters of the model.

**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 $\boldsymbol{h}_m$ must be computed from the previous hidden vector $\boldsymbol{h}_{m-1}$, 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. In contrast, **convolutional neural nets** predict each label $y_m$ from a set of matrix operations on the neighboring word embeddings, $\boldsymbol{x}_{m-k:m+k}$. 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 § 2.4.

## 6.6 *Unsupervised sequence labeling

In unsupervised sequence labeling, we want to induce a Hidden Markov Model from a corpus of unannotated text $\boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}, \ldots, \boldsymbol{w}^{(N)}$. 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.

Unsupervised learning in hidden Markov models can be performed using the **Baum-Welch algorithm**, which combines forward-backward with expectation-maximization (EM). In the M-step, we compute the HMM parameters from expected counts:

$$\Pr(W = i \mid Y = k) = \phi_{k,i} = \frac{E[\text{count}(W = i, Y = k)]}{E[\text{count}(Y = k)]}$$
$$\Pr(Y_m = k \mid Y_{m-1} = k') = \lambda_{k',k} = \frac{E[\text{count}(Y_m = k, Y_{m-1} = k')]}{E[\text{count}(Y_{m-1} = k')]}$$

The expected counts are computed in the E-step, using the forward and backward variables as defined in Equation 6.101 and Equation 6.104. Because we are working in a hidden Markov model, we define the potentials as,

$$\psi_m(k, k') = \text{p}_E(w_m \mid Y_m = k; \boldsymbol{\phi}) \times \text{p}_T(Y_m = k \mid Y_{m-1} = k'; \lambda). \quad [6.115]$$

The expected counts are then,

$$E[\text{count}(W = i, Y = k)] = \sum_{m=1}^{M} \Pr(Y_m = k \mid \boldsymbol{w}_{1:M})\delta(W_m = i) \quad [6.116]$$

$$= \sum_{m=1}^{M} \frac{\Pr(Y_m = k, \boldsymbol{w}_{1:m})\text{p}(\boldsymbol{w}_{m+1:M} \mid Y_m = k)}{\text{p}(\boldsymbol{w}_{1:M})}\delta(W_m = i) \quad [6.117]$$

$$= \frac{1}{\alpha_{M+1}(\blacklozenge)} \sum_{m=1}^{M} \alpha_m(k)\beta_m(k)\delta(W_m = i) \quad [6.118]$$

We use the chain rule to separate $\boldsymbol{w}_{1:m}$ and $\boldsymbol{w}_{m+1:M}$, and then use the definitions of the forward and backward variables. In the final step, we normalize by $\text{p}(\boldsymbol{w}_{1:M}) = \alpha_{M+1}(\blacklozenge) = \beta_0(\lozenge)$.

The expected transition counts can be computed in a similar manner:

$$E[\text{count}(Y_m = k, Y_{m-1} = k')] = \sum_{m=1}^{M} \Pr(Y_m = k, Y_{m-1} = k' \mid \boldsymbol{w}_{1:M}) \quad [6.119]$$

$$\propto \sum_{m=1}^{M} \text{p}(Y_{m-1} = k', \boldsymbol{w}_{1:m-1})\text{p}(w_{m+1:M} \mid Y_m = k) \times \text{p}(w_m, Y_m = k \mid Y_{m-1} = k') \quad [6.120]$$

$$= \sum_{m=1}^{M} \text{p}(Y_{m-1} = k', \boldsymbol{w}_{1:m-1})\text{p}(w_{m+1:M} \mid Y_m = k) \times \text{p}(w_m \mid Y_m = k)\text{p}(Y_m = k \mid Y_{m-1} = k') \quad [6.121]$$

$$= \sum_{m=1}^{M} \alpha_{m-1}(k')\beta_m(k)\phi_{k,w_m}\lambda_{k' \to k}. \quad [6.122]$$

Again, we use the chain rule to separate out $\boldsymbol{w}_{1:m-1}$ and $\boldsymbol{w}_{m+1:M}$, and use the definitions of the forward and backward variables. The final computation also includes the parameters $\phi$ and $\lambda$, which govern (respectively) the emission and transition properties between $w_m, y_m$, and $y_{m-1}$. Note that the derivation only shows how to compute this to a constant of proportionality; we would divide by $\text{p}(\boldsymbol{w}_{1:M})$ to go from the joint probability $\text{p}(Y_{m-1} = k', Y_m = k, \boldsymbol{w}_{1:M})$ to the desired conditional $\Pr(Y_{m-1} = k', Y_m = k \mid \boldsymbol{w}_{1:M})$.

### 6.6.1 Linear dynamical systems

The forward-backward algorithm can be viewed as Bayesian state estimation in a discrete state space. In a continuous state space, $y_m \in \mathbb{R}$, the equivalent algorithm is the **Kalman**

**Smoother**. It also computes marginals $\text{p}(y_m \mid \boldsymbol{x}_{1:M})$, using a similar two-step algorithm of forward and backward passes. Instead of computing a trellis of values at each step, we would compute a probability density function $q_{y_m}(y_m; \mu_m, \Sigma_m)$, characterized by a mean $\mu_m$ and a covariance $\Sigma_m$ around the latent state. Connections between the Kalman Smoother and the forward-backward algorithm are elucidated by Minka (1999) and Murphy (2012).

### 6.6.2 Alternative unsupervised learning methods

As noted in § 4.5, expectation-maximization is just one of many techniques for structure induction. One alternative is to use a family of randomized algorithms called **Markov Chain Monte Carlo (MCMC)**. In these algorithms, we compute a marginal distribution over the latent variable $\boldsymbol{y}$ **empirically**, by drawing random samples. The randomness explains the "Monte Carlo" part of the name; typically, we employ a Markov Chain sampling procedure, meaning that each sample is drawn from a distribution that depends only on the previous sample (and not on the entire sampling history). A simple MCMC algorithm is **Gibbs Sampling**, in which we iteratively sample each $y_m$ conditioned on all the others (Finkel et al., 2005):

$$\text{p}(y_m \mid \boldsymbol{y}_{-m}, \boldsymbol{w}_{1:M}) \propto \text{p}(w_m \mid y_m)\text{p}(y_m \mid \boldsymbol{y}_{-m}). \quad [6.123]$$

Gibbs Sampling has been applied to unsupervised part-of-speech tagging by Goldwater and Griffiths (2007). *Beam sampling* is a more sophisticated sampling algorithm, which randomly draws entire sequences $\boldsymbol{y}_{1:M}$, rather than individual tags $y_m$; this algorithm was applied to unsupervised part-of-speech tagging by Van Gael et al. (2009).

EM is guaranteed to find only a local optimum; MCMC algorithms will converge to the true posterior distribution $\text{p}(\boldsymbol{y}_{1:M} \mid \boldsymbol{w}_{1:M})$, but this is only guaranteed in the limit of infinite samples. Recent work has explored the use of **spectral learning** for latent variable models, which use matrix and tensor decompositions to provide guaranteed convergence under mild assumptions (Song et al., 2010; Hsu et al., 2012).

## Exercises

1. Consider the garden path sentence, *The old man the boat.* Given word-tag and tag-tag features, what inequality in the weights must hold for the correct tag sequence to outscore the garden path tag sequence for this example?
2. Sketch out an algorithm for a variant of Viterbi that returns the top-$n$ label sequences. What is the time and space complexity of this algorithm?
3. Show how to compute the marginal probability $\Pr(Y_{m-2} = k, Y_m = k')$, in terms of the forwards and backward variables, and the potentials $\psi_m$.

4. more tk

# Chapter 7

# Applications of sequence labeling

Sequence labeling has applications throughout natural language processing. This chapter focuses on the classical applications of part-of-speech tagging, morpho-syntactic attribute tagging, named entity recognition, and tokenization. It also touches briefly on two applications to interactive settings: dialogue act recognition and the detection of code-switching points between languages.

## 7.1 Part-of-speech tagging

The **syntax** of a language is the collection of principles under which sequences of words are judged to be grammatically acceptable by fluent speakers. One of the most basic syntactic concepts is the **part-of-speech** (POS), which refers to the syntactic role of each word in a sentence. We have already referred to this concept informally in the previous chapter, and you may have some intuitions from your own study of English. For example, in the sentence *We like vegetarian sandwiches*, you may already know that *we* and *sandwiches* are nouns, *like* is a verb, and *vegetarian* is an adjective. These labels depend on the context in which the word appears: in *she eats like a vegetarian*, the word *like* is a preposition, and the word *vegetarian* is a noun.

Parts-of-speech can help to disentangle or explain various linguistic problems. Recall Chomsky's proposed distinction in chapter 5:

(7.1) Colorless green ideas sleep furiously.

(7.2) *Ideas colorless furiously green sleep.

Why is the first grammatical and the second not? One explanation is that the first example contains part-of-speech transitions that are typical in English: adjective to adjective, adjective to noun, noun to verb, and verb to adverb. In contrast, the second sentence contains transitions that are unusual: noun to adjective and adjective to verb. The ambiguity in

sentences like *teacher strikes idle children* can also be explained in terms of parts of speech: in the interpretation that was likely intended, *strikes* is a noun and *idle* is a verb; in the alternative explanation, *strikes* is a verb and *idle* is an adjective.

Part-of-speech tagging is often taken as a early step in a natural language processing pipeline. Indeed, parts-of-speech provide features that can be useful for many of the tasks that we will encounter later, such as parsing, coreference resolution, and relation extraction.

### 7.1.1 Parts-of-Speech

The **Universal Dependencies** project (UD) is an effort to create syntactically-annotated corpora across many languages, using a single annotation standard (Nivre et al., 2016). As part of this effort, they have designed a part-of-speech tag inventory, which is meant to capture word classes across as many languages as possible.[1] This section describes that inventory, giving rough definitions for each of tags, along with supporting examples.

Part-of-speech tags are **morphosyntactic**, rather than **semantic**, categories. This means that they describe words in terms of how they pattern together and how they are internally constructed (e.g., what suffixes and prefixes they include). For example, you may think of a noun as referring to objects or concepts, and verbs as referring to actions or events. But events can also be nouns — for example, *the **running** of the bulls*, where the determiner *the* is a strong clue that *running* is acting as a noun.

#### The Universal Dependency part-of-speech tagset

The UD tagset is broken up into three groups: open class tags, closed class tags, and "other".

**Open class tags** Nearly all languages contain nouns, verbs, adjectives, and adverbs.[2] These are all **open word classes**, because new words are constantly being added to them. The UD tagset includes two other tags that are open classes: proper nouns and interjections.

- **Nouns** (UD tag: NOUN) tend to describe entities and concepts, e.g.,

  (7.3) **Toes** are scarce among veteran **blubber men**.

  In English, nouns tend to follow determiners and adjectives, and can play the subject role in the sentence. They can be marked for the plural number by an *-s* suffix.

[1] The 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.

[2] One well-known exception is Korean, which some linguists argue does not have adjectives Kim (2002).

- **Proper nouns** (PROPN) are tokens in names, which uniquely specify a given entity,

  (7.4) "**Moby Dick**?" shouted **Ahab**.

- **Verbs** (VERB) describe actions or events

  (7.5) "Moby Dick?" **shouted** Ahab.

  (7.6) **Shall** we **keep chasing** this murderous fish?

  English verbs tend to come in between the subject and some number of direct objects, depending on the verb. They can be marked for **tense** and **aspect** using suffixes such as *-ed* and *-ing*. (These suffixes are an example of **inflectional morphology**, which is discussed in more detail in § 8.1.4.)

- **Adjectives** (ADJ) describe properties of entities,

  (7.7) Shall we keep chasing this **murderous** fish?

  (7.8) Toes are **scarce** among **veteran** blubber men.

  In the second example, *scarce* is a predicative adjective, linked to the subject by the **copula verb** *are*. In contrast, *murderous* and *veteran* are attribute adjectives, modifying the noun phrase in which they are embedded.

- **Adverbs** (ADV) describe properties of events, and may also modify adjectives or other adverbs:

  (7.9) It is not down on any map; true places **never** are.

  (7.10) . . . **treacherously** hidden beneath the loveliest tints of azure

  (7.11) Not drowned **entirely**, though.

- **Interjections** (INTJ) are used in exclamations, e.g.,

  (7.12) **Aye aye**! it was that accursed white whale that razed me.

**Closed class tags** Closed word classes rarely receive new members. They are sometimes referred to as **function words** — as opposed to **content words** — as they have little lexical meaning of their own, but rather, help to organize the components of the sentence.

- **Adpositions** (ADP) describe the relationship between a complement (usually a noun phrase) and another unit in the sentence, typically a noun or verb phrase.

  (7.13) Toes are scarce **among** veteran blubber men.

(7.14) It is not **down on** any map.

(7.15) Give not thyself **up** then.

As the examples show, English generally uses prepositions, which are adpositions that appear before their complement. (An exception is *ago*, as in, *we met three days **ago***). Postpositions are used in other languages, such as Japanese and Turkish.

- **Auxiliary verbs** (AUX) are a closed class of verbs that add information such as tense, aspect, person, and number.

  (7.16) **Shall** we keep chasing this murderous fish?

  (7.17) What the white whale was to Ahab, **has been** hinted.

  (7.18) Ahab **must** use tools

  (7.19) Meditation and water **are** wedded forever.

  (7.20) Toes **are** scarce among veteran blubber men.

  As noted above, the final example is a copula verb, which is also tagged as an auxiliary in the UD corpus.

- **Coordinating conjunctions** (CCONJ) express relationships between two words or phrases, which play a parallel role:

  (7.21) Meditation **and** water are wedded forever.

- **Subordinating conjunctions** (SCONJ) link two elements, making one syntactically subordinate to the other:

  (7.22) There is wisdom **that** is woe.

- **Pronouns** (PRON) are words that substitute for nouns or noun phrases when the meaning is clear from context.

  (7.23) Be **it what it** will, **I** 'll go to **it** laughing.

  (7.24) **I** try all things, **I** achieve **what I** can.

  The example includes the personal pronouns *I* and *it*, as well as the relative pronoun *what*. Other pronouns include *myself*, *somebody*, and *nothing*.

- **Determiners** (DET) provide additional information about the nouns or noun phrases that they modify:

  (7.25) What **the** white whale was to Ahab, has been hinted.

  (7.26) It is not down on **any** map.

(7.27) I try **all** things ...

(7.28) Shall we keep chasing **this** murderous fish?

Determiners include articles (*the*), possessive determiners (*their*), demonstratives (*this murderous fish*), and quantifiers (*any map*).

- **Numerals** (NUM) are an infinite but closed class, which includes integers, fractions, and decimals, regardless of whether spelled out or written in numerical form.

  (7.29) How then can this **one** small heart beat

  (7.30) I am going to put him down for the **three hundredth**

- **Particles** (PART) are a catch-all of function words that combine with other words or phrases, but do not meet the conditions of the other tags. In English, this includes the infinitival *to*, the possessive marker, and negation.

  (7.31) Better **to** sleep with a sober cannibal than a drunk Christian.

  (7.32) So man **'s** insanity is heaven **'s** sense

  (7.33) It is **not** down on any map

  As the first example shows, the possessive marker is not considered part of the same token as the word that it modifies, so that *man's* is split into two tokens. (Tokenization is described in more detail in § 7.4.) Non-English examples of particles include question markers, such as

  In other tagsets, particularly those designed for English like the Penn Treebank, verbal particles such as *give **up*** and *fuck **off*** are tagged as particles. However, in the UD corpus, these are tagged as adpositions.

**Other** Remaining UD tags include punctuation (PUN) and symbols (SYM). Punctuation is purely structural — e.g., commas, periods, colons — while symbols can carry content of their own. Examples of symbols include dollar and percentage symbols, mathematical operators, emoticons, emojis, and internet addresses. A final catch-all tag is X, which is used for words that cannot be assigned another part-of-speech category. The X tag is also used in cases of **code switching** (between languages), described in § 7.5.

#### Other tagsets

Prior to the Universal Dependency treebank, part-of-speech tagging was performed using language-specific tagsets. The dominant tagset for English is the Penn Treebank (PTB). The PTB tagset includes 45 tags — much more granularity than the UD tagset. This granularity is reflected in distinctions between singular and plural nouns, various verb tenses

and aspects, possessive and non-possessive pronouns, comparative and superlative adjectives and adverbs (e.g., *faster*, *fastest*), and so on. The Brown corpus includes a tagset that is even more detailed, with 87 tags Francis (1964), including special tags for individual auxiliary verbs such as *be*, *do*, and *have*.

Different languages make different distinctions, and so the PTB and Brown tagsets are not appropriate for a language such as Chinese, which does not mark the verb tense (Xia, 2000); nor for Spanish, which marks every combination of person and number in the verb ending; nor for German, which marks the case of each noun phrase. Each of these languages requires more detail than English in some areas of the tagset, and less in other areas. The strategy of the Universal Dependency corpus is to design a coarse-grained tagset to be used across all languages, and then to additionally annotate language-specific **morphosyntactic attributes**, such as number, tense, and case. The attribute tagging task is described in more detail in § 7.2.

Finally, social media such as Twitter have been shown to require tagsets of their own (Gimpel et al., 2011). Such corpora contain some tokens that are simply not equivalent to anything encountered in a typical written corpus: e.g., emoticons, URLs, and hashtags. Social media also includes dialectal words like *gonna* (going to, e.g. *We gonna be fine*) and *Ima* (I'm going to, e.g., *Ima tell you one more time*), which can be analyzed either as non-standard orthography (making tokenization impossible), or as lexical items in their own right. In either case, it is clear that existing tags like NOUN and VERB cannot handle cases like *Ima*, which combine aspects of the noun and verb. Gimpel et al. (2011) therefore propose a new set of tags to deal with these cases.

### 7.1.2 Accurate part-of-speech tagging

Part-of-speech tagging is the problem of selecting the correct tag for each word in a sentence. Success is typically measured by accuracy on an annotated test set, which is simply the fraction of tokens that were tagged correctly.

#### Baselines

A simple baseline for part-of-speech tagging is to choose the most common tag for each word. For example, in the Universal Dependency treebank, the word *talk* appears 96 times, and 85 of those times it is labeled as a verb: therefore, this baseline will always predict verb when we see it in the test set. For words that do not appear in the training corpus, the baseline simply guesses the most common tag overall, which is noun. In the Penn Treebank, this simple baseline obtains accuracy above 92%. A more rigorous evaluation is the accuracy on **out-of-vocabulary words**, which are not seen in the training data. Tagging the words correctly requires attention to the context or the word's internal structure.

**Contemporary approaches**

Conditional random fields and structured perceptron perform at or near the state-of-the-art for part-of-speech tagging in English. For example, (Collins, 2002) achieved 97.1% accuracy on the Penn Treebank, using a structured perceptron, using the following base features (originally introduced by Ratnaparkhi (1996)):

- current word, $w_m$
- previous words, $w_{m-1}, w_{m-2}$
- next words, $w_{m+1}, w_{m+2}$
- previous tag, $y_{m-1}$
- previous two tags, $(y_{m-1}, y_{m-2})$
- for rare words:
  - first $k$ characters, up to $k = 4$
  - last $k$ characters, up to $k = 4$
  - whether $w_m$ contains a number, uppercase character, or hyphen.

Similar results for the PTB data have been achieved using conditional random fields (CRFs; Toutanova et al., 2003).

More recent work has demonstrated the power of neural sequence models, such as the **long short-term memory (LSTM)** (§ 6.5.4). Plank et al. (2016) apply a CRF and a bidirectional LSTM to twenty-two languages in the UD corpus, achieving an average accuracy of 94.3% for the CRF, and 96.5% with the bi-LSTM. Their model employs three types of embeddings: learned word embeddings, which are updated during training; pre-trained word embeddings, which are never updated, but which help to tag out-of-vocabulary words; and character-based embeddings. These character-based embeddings are computed by running an LSTM on the individual characters in each word, thereby capturing common orthographic patterns such as prefixes, suffixes, and capitalization. Extensive evaluations show that these additional embeddings are crucial to their model's success.

## 7.2 Morphosyntactic Attributes

The Universal Dependency (UD) tagset provides a coarse-grained description of the syntactic role of each word in a sentence, using a set of tags that generalizes across languages. However, there is considerably more to say about a word than whether it is a noun or a verb: in English, verbs are distinguish by features such tense and aspect, nouns by number, adjectives by degree, and so on. Unfortunately, these features are language-specific: other languages distinguish other features, such as **case** (the role of the noun with respect to the action of the sentence, which is marked in many languages, such as Latin and Ger-

| word | PTB tag | UD tag | UD attributes |
|---|---|---|---|
| *The* | DT | DET | DEFINITE=DEF |
| | | | PRONTYPE=ART |
| *German* | JJ | ADJ | DEGREE=POS |
| *Expressionist* | NN | NOUN | NUMBER=SING |
| *movement* | NN | NOUN | NUMBER=SING |
| *was* | VBD | AUX | MOOD=IND |
| | | | NUMBER=SING |
| | | | PERSON=3 |
| | | | TENSE=PAST |
| | | | VERBFORM=FIN |
| *destroyed* | VBN | VERB | TENSE=PAST |
| | | | VERBFORM=PART |
| | | | VOICE=PASS |
| *as* | IN | ADP | |
| *a* | DT | DET | DEFINITE=IND |
| | | | PRONTYPE=ART |
| *result* | NN | NOUN | NUMBER=SING |
| . | . | PUNCT | |

Figure 7.1: UD and PTB part-of-speech tags, and UD morphosyntactic attributes. Example selected from the UD 1.4 English corpus.

man[3] and **evidentiality** (the source of information for the speaker's statement, which is marked in languages such as Turkish). In the UD corpus, these attributes are annotated as feature-value pairs for each token.[4]

An example is shown in Figure 7.1. The determiner *the* is marked with two attributes: PRONTYPE=ART, which indicates that it is an **article** (as opposed to another type of determiner or pronominal modifier), and DEFINITE=DEF, which indicates that it is a **definite article** (referring to a specific, known entity). The verbs are each marked with several attributes. The auxiliary verb *was* is third-person, singular, past tense, finite (conjugated), and indicative (describing an event that has happened or is currently happenings); the main verb *destroyed* is in participle form (so there is no additional person and number information), past tense, and passive voice. Some, but not all, of these distinctions are reflected in the PTB tags VBD (past-tense verb) and VBN past participle.

[3]Case is marked in English for some personal pronouns, e.g., ***She*** *saw* ***her***, *They* saw *them*.

[4]The 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).

While there are thousands of papers on part-of-speech tagging, there is comparatively little work on automatically labeling morphosyntactic attributes. Faruqui et al. (2016) train a support vector machine classification model, using a minimal feature set that includes the word itself, its prefixes and suffixes, and type-level information listing all possible morphosyntactic attributes for each word and its neighbors. Mueller et al. (2013) use a conditional random field (CRF), in which the tag space consists of all observed combinations of morphosyntactic attributes (e.g., the tag would be DEF+ART for the word *the* in Figure 7.1). This massive tag space is managed by decomposing the feature space over individual attributes, and pruning paths through the trellis. More recent work has employed bidirectional LSTM sequence models. For example, Pinter et al. (2017) train a bidirectional LSTM sequence model. The input layer and hidden vectors in the LSTM are shared across attributes, but each attribute has its own output layer, culminating in a softmax over all attribute values, e.g. $y_t^{\text{NUMBER}} \in \{\text{SING}, \text{PLURAL}, \ldots\}$. They find that character-level information is crucial, especially when the amount of labeled data is limited.

Evaluation is performed by first computing recall and precision for each attribute. These scores can then be averaged at either the type or token level to obtain micro- or macro-$F$-measure. Pinter et al. (2017) evaluate on 23 languages in the UD treebank, reporting a median micro-$F$-measure of 0.95. Performance is strongly correlated with the size of the labeled dataset for each language, with a few outliers: for example, Chinese is particularly difficult, because although the dataset is relatively large ($10^5$ tokens in UD 1.4), only 6% of tokens have any attributes, offering few useful labeled instances.

## 7.3 Named entity recognition

A core problem in information extraction is to recognize and extract mentions of **named entities** in text. In news documents, the core entity types are people, locations, and organizations; more recently, the task has been extended to include amounts of money, percentages, dates, and times. In (7.34) in Figure 7.2, the named entities include: *The U.S. Army*, an organization; *Atlanta*, a location; and *May 14, 1864*, a date. Named entity recognition is also a key task in **biomedical natural language processing**, with entity types including proteins, DNA, RNA, and cell lines (e.g., Collier et al., 2000; Ohta et al., 2002). Figure 7.2 shows an example from the GENIA corpus.

A standard approach is to tagging named entity spans is to use discriminative sequence labeling methods such as conditional random fields and structured perceptron. As described in chapter 6, these methods use the Viterbi algorithm to search over all possible label sequences, while scoring each sequence using a feature function that decomposes across adjacent tags. Named entity recognition is formulated as a tagging problem by assinging each word token to a tag from a tagset. However, there is a major difference from part-of-speech tagging: in NER we need to recover **spans** of tokens, such as *The*

(7.34) *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

(7.35) *Number of glucocorticoid receptors in lymphocytes and ...*
O O B-PROTEIN I-PROTEIN O B-CELLTYPE O ...

Figure 7.2: BIO notation for named entity recognition. Example (7.35) is drawn from the GENIA corpus of biomedical documents (Ohta et al., 2002).

*United States Army*. To do this, the tagset must distinguish tokens that are at the **b**eginning of a span from tokens that are **i**nside a span.

This is accomplished by the **BIO notation**, shown in Figure 7.2. Each token at the beginning of a name span is labeled with a B- prefix; each token within a name span is labeled with an I- prefix. Tokens that are not parts of name spans are labeled as O. From this representation, it is unambiguous to recover the entity name spans within a labeled text. Another advantage is from the perspective of learning: tokens at the beginning of name spans may have different properties than tokens within the name, and the learner can exploit this. This insight can be taken even further, with special labels for the **l**ast tokens of a name span, and for **u**nique tokens in name spans, such as *Atlanta* in the example in Figure 7.2. This is called BILOU notation, and has been shown to yield improvements in supervised named entity recognition (Ratinov and Roth, 2009).

**Feature-based sequence labeling** Named entity recognition was an early application of conditional random fields (McCallum and Li, 2003). The use of Viterbi decoding restricts the feature function $\boldsymbol{f}(\boldsymbol{w}, \boldsymbol{y}) = \sum_m \boldsymbol{f}(\boldsymbol{w}, y_m, y_{m-1}, m)$, 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 to capture capitalization. As an example, base features for the word *Army* in the example in (7.34) include:

$\langle$CURR-WORD:*Army*, PREV-WORD:*U.S.*, NEXT-WORD:*captured*, PREFIX-1:*A-*,
PREFIX-2:*Ar-*, SUFFIX-1:*-y*, SUFFIX-2:*-my*, SHAPE:*Xxxx*$\rangle$

Another source of features is to use **gazzeteers**: lists of known entity names. For example, the U.S. Social Security Administration provides a list of tens of thousands of given names — more than could be observed in any reasonable annotated corpus. Tokens or spans that match an entry in a gazetteer can receive special features; this provides a way to incorporate hand-crafted resources such as name lists in a learning-driven framework.

**Neural sequence labeling for NER** Current research has emphasized neural sequence labeling, using similar LSTM models to those employed in part-of-speech tagging (Ham-

[todo: import zh-example successfully]

Figure 7.3: Example of Chinese tokenization (Sproat et al., 1996)

merton, 2003; Huang et al., 2015; Lample et al., 2016). The bidirectional LSTM-CRF (Figure 6.3 in § 6.5.4) does particularly well on this task, due to its ability to model tag-to-tag dependencies. However, Strubell et al. (2017) show that **convolutional neural networks** can be equally accurate, with significant improvement in speed due to the efficiency of implementing ConvNets on **graphics processing units (GPUs)**. The key innovation in this work was the use of **dilated convolutions**, which are described in more detail in § 2.4.

## 7.4 Tokenization

A basic problem for text analysis, first discussed in § 3.3.1, is to break the text into a sequence of discrete tokens. For alphabetic languages such as English, deterministic scripts suffice to achieve accurate tokenization. However, in logographic writing systems such as Chinese script, words are typically composed of a small number of characters, without space in between (Figure 7.3). One approach is to match character sequences against a known dictionary (e.g., Sproat et al., 1996), using additional statistical information about word frequency. However, no dictionary is completely comprehensive, and dictionary-based approaches can struggle with such out-of-vocabulary words.

Chinese tokenization has therefore been approached as a supervised sequence labeling problem. Xue et al. (2003) train a logistic regression classifier to make independent segmentation decisions while moving a sliding window across the document. A set of rules is then used to convert these individual classification decisions into an overall tokenization of the input. However, these individual decisions may be globally suboptimal, motivating a structure prediction approach. Peng et al. (2004) train a conditional random field model to predict labels of START or NONSTART on each character. More recent work has employed neural network architectures. For example, Chen et al. (2015) use an LSTM-CRF architecture, as described in § 6.5.4: they construct a trellis, in which each tag is scored according to the hidden state of an LSTM, and tag-tag transitions are scored according to learned transition weights. The best-scoring segmentation is then computed by the Viterbi algorithm.

## 7.5 Code switching

[todo: needs work] Multilingual speakers and writers do not restrict themselves to a single language; rather, they may switch between languages, in a phenomenon known as **code switching** (Auer, 2013; Poplack, 1980). Code switching is particularly common in

online social media, as in the following extract from Justin Trudeau's website:[5]

(7.36) *Although everything written on this site est disponible en anglais*
is available in English
*and in French, my personal videos seront bilingues*
will be bilingual

Accurately analyzing such texts requires first determining which languages are being used. Furthermore, quantitative analysis of code switching can provide insights on the 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) attack the problem of code switching between Turkish and Dutch, using a conditional random field.

The problem of identifying code switching points 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 belonging to one of the three languages, as a mix of multiple languages, as "universal" (e.g., symbols, numbers, emoticons), or as undefined.

## 7.6 Dialogue acts

The sequence labeling problems that we have discussed so far have been over sequences of word tokens or characters (in the case of tokenization). However, sequence labeling can also be performed over higher-level units, such as **utterances**. **Dialogue acts** are labels over utterances in a dialogue, corresponding roughly to the speaker's intention — its **illocutionary force** (Austin, 1962). For example, an utterance may state a proposition (*it is not down on any map*), pose a question (*shall we keep chasing this murderous fish?*), or provide a response (*aye aye!*). Stolcke et al. (2000) describe how a set of 42 dialogue acts were annotated for the 1,155 conversations in the Switchboard corpus (Godfrey et al., 1992).[6]

[5] As quoted in `http://blogues.lapresse.ca/lagace/2008/09/08/justin-trudeau-really-parfait-bilingue/`, accessed August 21, 2017.

[6] Dialogue 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

| Speaker | Dialogue Act | Utterance |
|---|---|---|
| A | YES-NO-QUESTION | *So do you go college right now?* |
| A | ABANDONED | *Are yo-* |
| B | YES-ANSWER | *Yeah,* |
| B | STATEMENT | *It's my last year [laughter].* |
| A | DECLARATIVE-QUESTION | *You're a, so you're a senior now.* |
| B | YES-ANSWER | *Yeah,* |
| B | STATEMENT | *I'm working on my projects trying to graduate [laughter]* |
| A | APPRECIATION | *Oh, good for you.* |
| B | BACKCHANNEL | *Yeah.* |

Figure 7.4: Example of dialogue act labeling (Stolcke et al., 2000).

An example is shown in Figure 7.4. The annotation is performed over UTTERANCES, with the possibility of multiple segments per **conversational turn** (in rare cases such as interruptions, an utterance may split over multiple turns). Some segments are clauses (e.g., *So do you go to college right now?*), while others are single words (e.g., *yeah*). Stolcke et al. (2000) report that hidden Markov models (HMMs) achieve 96% accuracy on supervised utterance segmentation. The labels themselves reflect the conversational goals of the speaker: the utterance *yeah* functions as an answer in response to the question *you're a senior now*, but as a **backchannel** (demonstrating comprehension and encouraging the other speaker to continue talking) in the final line of the excerpt.

For task of dialogue act labeling, Stolcke et al. (2000) apply a hidden Markov model. The generative probability $\text{p}(\boldsymbol{w}_m \mid y_m)$ must generate the entire sequence of words in the utterance, and it is modeled as a trigram language model (§ 5.1). Stolcke et al. (2000) also account for acoustic features, which capture the **prosody** of each utterance — for example, tonal and rhythmic properties of speech, which can be used to distinguish dialogue acts such as questions and answers. These features are handled with an additional emission distribution, $\text{p}(\boldsymbol{a}_m \mid y_m)$, which is modeled with a probabilistic decision tree (Murphy, 2012). They find that while acoustic features yield small improvements overall, they play an important role in distinguish questions from statements, and agreements from backchannels.

Recurrent neural architectures for dialogue act labeling have been proposed by Kalchbrenner and Blunsom (2013) and Ji et al. (2016), with strong empirical results. Both models are recurrent at the utterance level, so that each complete utterance updates a hidden state. The recurrent-convolutional network of Kalchbrenner and Blunsom (2013) uses convolution to obtain a representation of each individual utterance, while Ji et al. (2016) use a second level of recurrence, over individual words. This enables their method to also func-

forums.

tion as a language model, giving probabilities over sequences of words in a document.

# Chapter 8

# Formal language theory

We have now seen methods for learning to label individual words, vectors of word counts, and sequences of words; we will soon proceed to more complex structural transformations. Most of these techniques could apply to counts or sequences from any discrete vocabulary; there is nothing fundamentally linguistic about, say, a hidden Markov model. This raises an embarrassingly basic question that this text has not yet considered: what is a language?

This chapter will take the perspective of **formal language theory**, in which a **language** is defined as a set of **strings**, each of which is a sequence of elements from a finite alphabet. For interesting languages, there are an infinite number of strings that are in the language, and an infinite number of strings that are not. For example:

- the set of all even-length sequences from the alphabet $\{a, b\}$, e.g., $\{\varnothing, aa, ab, ba, bb, aaaa, aaab, \ldots\}$;
- the set of all sequences from the alphabet $\{a, b\}$ that contain $aaa$ as a substring, e.g., $\{aaa, aaaa, baaa, aaab, \ldots\}$;
- the set of all sequences of English words (drawn from a finite dictionary) that contain at least one verb (a finite subset of the dictionary);
- the `python` programming language.

Formal language theory has defined various classes of languages and their computational properties. In particular, theorists have delineated classes of languages based on the computational complexity of solving the **membership problem** — determining whether a string is in a language. The chapter will focus on three classes of formal languages: regular, context-free, and "mildly" context-sensitive languages.

A key insight of 20th century linguistics is that formal language theory can be usefully applied to natural languages such as English, by designing formal languages that capture as many properties of the natural language as possible. For many such formalisms, a

useful linguistic analysis comes as a byproduct of solving the membership problem. Furthermore, the membership problem can be generalized to the problems of **scoring** strings for their acceptability (as in language modeling), and of **transducing** one string into another.

## 8.1 Regular languages

Sooner or later, most computer scientists will write a **regular expression**. If you have, then you have defined a **regular language**, which is any language that can be defined by a regular expression. Formally, a regular expression can include the following elements:

- A **literal character** drawn from some finite alphabet $\Sigma$.
- The **empty string** $\epsilon$.
- The concatenation of two regular expressions $RS$, where $R$ and $S$ are both regular expressions. The resulting expression accepts a string that can be decomposed $x = yz$, where $y$ is accepted by $R$ and $z$ is accepted by $S$.
- The alternation $R \mid S$, where $R$ and $S$ are both regular expressions. The resulting expression accepts a string $x$ if it is accepted by $R$ or it is accepted by $S$.
- The **Kleene star** $R^*$, which accepts any string $x$ that can be decomposed into a sequence of strings which are all accepted by $R$.
- Parenthesization $(R)$, which is used to limit the scope of the concatenation, alternation, and Kleene star operators.

Here are some example regular expressions for some of the languages described above:

- The set of all even length strings on the alphabet $\{a, b\}$: $((aa)|(ab)|(ba)|(bb))^*$
- The set of all sequences of the alphabet $\{a, b\}$ that contain $aaa$ as a substring: $(a|b)^*aaa(a|b)^*$
- The set of all sequences of English words that contain at least one verb: $W^*VW^*$, where $W$ is an alternation between all words in the dictionary, and $V$ is an alternation between all verbs ($V \subseteq W$).

We do not include a regular expression for the Python programming language, because this language is not regular — there is no regular expression that can capture its syntax. We will discuss why towards the end of this section.

Regular languages are **closed** under union, intersection, and concatenation. This means, for example, that if two languages $L_1$ and $L_2$ are regular, then so are the languages $L_1 \cup L_2$, $L_1 \cap L_2$, and the language of strings that can be decomposed as $s = tu$, with $s \in L_1$ and $t \in L_2$. Regular languages are also closed under negation: if $L$ is regular, then so is the language $\overline{L} = \{s \notin L\}$.

### 8.1.1 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 algorithms for 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 $M = \langle Q, \Sigma, q_0, F, \delta \rangle$, consisting of:

- a finite alphabet $\Sigma$ of input symbols;
- a finite set of **states** $Q = \{q_0, q_1, \ldots, q_n\}$;
- 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 $\epsilon$) to a **set** of possible resulting states.

A **path** in $M$ is a sequence of transitions, $\pi = t_1, t_2, \ldots, t_N$, where each $t_i$ traverses an arc in the transition function $\delta$. The finite-state acceptor $M$ accepts a string $\omega$ if there is a **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 $\omega$ is consumed.

#### Example

Consider the following FSA, $M_1$.

$$\Sigma = \{a, b\} \quad [8.1]$$
$$Q = \{q_0, q_1\} \quad [8.2]$$
$$F = \{q_1\} \quad [8.3]$$
$$\delta = \{\{(q_0, a) \rightarrow q_0\}, \{(q_0, b) \rightarrow q_1\}, \{(q_1, b) \rightarrow q_1\}\} \quad [8.4]$$

This FSA defines a language over an alphabet of two symbols, *a* and *b*. The transition function $\delta$ is written as a set of tuples: the tuple $\{(q_0, a) \rightarrow q_0\}$ says that if the machine is in state $q_0$ and reads symbol $a$, it stays in $q_0$. A graphical state diagram representation for $M_1$ is shown in Figure 8.1. Because each pair of initial state and symbol has at most one resulting state, $M_1$ is **deterministic**: each string $\omega$ induces at most one accepting path. Note that there are no transitions for the symbol $a$ in state $q_1$; if $a$ is encountered in $q_1$, then the acceptor is stuck, and the input string is rejected.

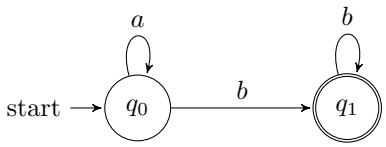



Figure 8.1: State diagram for the finite-state acceptor $M_1$.

What strings does $M_1$ accept? The start state is $q_0$, and we have to get to $q_1$, since this is the only final state. Any number of $a$ symbols can be consumed in $q_0$, but a $b$ symbol is required to transition to $q_1$. Once there, any number of $b$ symbols can be consumed, but an $a$ symbol cannot. So the regular expression corresponding to the language defined by $M_1$ is $a^*bb^*$.

**Computational properties of finite-state acceptors**

The key computational question for finite-state acceptors is: how fast can we determine whether a string is accepted? For deterministic FSAs, this computation can be performed by Dijkstra's algorithm, with time complexity $\mathcal{O}(V \log V + E)$, where $V$ is the number of vertices in the FSA, and $E$ is the number of edges (Cormen et al., 2009). Non-deterministic FSAs (NFSAs) can include multiple transitions from a given symbol and state. Any NSFA can be converted into a deterministic FSA, but the resulting automaton may have a number of states that is exponential in the number of size of the original NFSA (Mohri et al., 2002).

### 8.1.2 Morphology as a regular language

Many words have internal structure, such as prefixes and suffixes that shape their meaning. The study of word-internal structure is the domain of **morphology**, of which there are two main types:

- **Derivational morphology** describes the use of affixes to convert a word from one grammatical category to another (e.g., from the noun *grace* to the adjective *graceful*), or to change the meaning of the word (e.g., from *grace* to *disgrace*).
- **Inflectional morphology** describes the addition of details such as gender, number, person, and tense (e.g., the *-ed* suffix for past tense in English).

Morphology is a rich topic in linguistics, deserving of a course in its own right.[1] The focus here will be on the use of finite-state automata for morphological analysis. The

[1] A 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*. In semitic languages such as Arabic and Hebrew, morphology involves a template-based system, in which roots

current section deals with derivational morphology; inflectional morphology is discussed in § 8.1.4.

Suppose that we want to write a program that accepts only those words that are constructed in accordance with the rules of English derivational morphology:

(8.1) grace, graceful, gracefully, *gracelyful

(8.2) disgrace, *ungrace, disgraceful, disgracefully

(8.3) allure, *allureful, alluring, alluringly

(8.4) fairness, unfair, *disfair, fairly

(Recall that the asterisk indicates that a linguistic example is judged unacceptable by fluent speakers of a language.) These examples cover only a tiny corner of English derivational morphology, but a number of things stand out. The suffix *-ful* converts the nouns *grace* and *disgrace* into adjectives, and the suffix *-ly* converts adjectives into adverbs. These suffixes must be applied in the correct order, as shown by the unacceptability of **gracelyful*. The *-ful* suffix only works for some words, as shown by the use of *alluring* as the adjectival form of of *allure*. Other changes are made with prefixes, such as the derivation of *disgrace* from *grace*, which roughly corresponds to a negation; however, the negation of *fair* is performed with the *un-* prefix instead. Finally, while the first three examples suggest that the direction of derivation is noun → adjective → adverb, the example of *fair* suggests that the adjective can also be the base form, with the *-ness* suffix performing the conversion to a noun.

These examples lead to a computational question: can we build a computer program that accepts only well-formed English words, and rejects all others? This might at first seem trivial to solve with a brute-force attack: simply make a dictionary of all valid English words, as in a spell-checker. But such an approach fails to account for morphological **productivity** — the applicability of existing morphological rules to new words and names, such as *Trump* to *Trumpy* and *Trumpkin*, and *Clinton* to *Clintonian* and *Clintonite*. We need an approach that represents morphological rules explicitly, and for this we will try a finite state acceptor.

The dictionary approach can be implemented as a finite state acceptor, with the vocabulary $\Sigma$ equal to the vocabulary of English, and a transition from the start state to the accepting state for each word. But this would of course fail to generalize beyond the original vocabulary, and would not capture anything about the **morphotactic** rules that govern derivations from new words. The first step towards a more general approach is shown in Figure 8.2, which is the state diagram for a finite-state transducer in which the vocabulary

are triples of consonants (e.g., *ktb*), and words are created by adding consonants: ***k**a**t**a**b**a* (Arabic: he wrote), ***k**u**t**u**b*** (books), *ma**k**ta**b*** (desk). For more detail on morphology, see texts from Haspelmath and Sims (2013) and Lieber (2015).

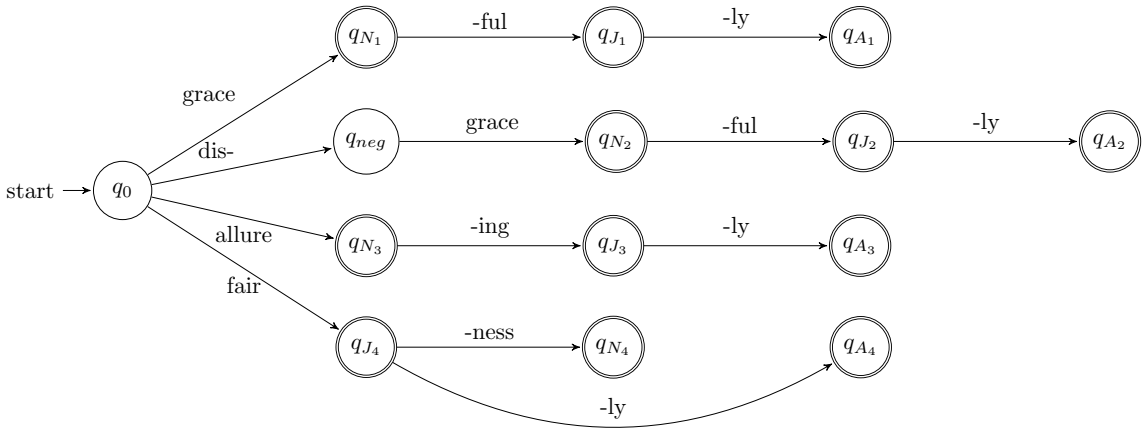



Figure 8.2: A finite-state acceptor for a fragment of English derivational morphology. Each path represents possible derivations from a single root form.

consists of **morphemes**, which include **stems** (e.g., *grace, allure*) and **affixes** (e.g., *dis-*, *-ing*, *-ly*). This finite-state acceptor consists of a set of paths leading away from the start state, with derivational affixes added along the path. Except for $q_{neg}$, the states on these paths are all final, so the FSA will accept *disgrace*, *disgraceful*, and *disgracefully*, but not *dis-*.

This FSA can be **minimized** to the form shown in Figure 8.3, which makes the generality of the finite-state approach more apparent. For example, the transition from $q_0$ to $q_{J_2}$ can be made to accept not only *fair* but any single-morpheme (**monomorphemic**) adjective that takes *-ness* and *-ly* as suffixes. In this way, the finite-state acceptor can easily be extended: as new word stems are added to the vocabulary, their derived forms will be accepted automatically. Of course, this FSA would still need to be extended considerably to cover even this small fragment of English morphology. As shown by cases like *music* $\rightarrow$ *musical*, *athlete* $\rightarrow$ *athletic*, English includes several classes of nouns, each with its own rules for derivation.

The FSAs shown in Figure 8.2 and 8.3 accept *allureing*, not *alluring*. This reflects a distinction between morphology — the question of which morphemes and in what order — and **orthography** — the question of how the morphemes are rendered in written language. Just as orthography requires dropping the *e* preceding the *-ing* suffix, **phonology** imposes a related set of constraints on how words are rendered in speech. As we will see soon, these issues are handled through **finite-state transducers**, which are finite-state automata that take inputs and produce outputs.

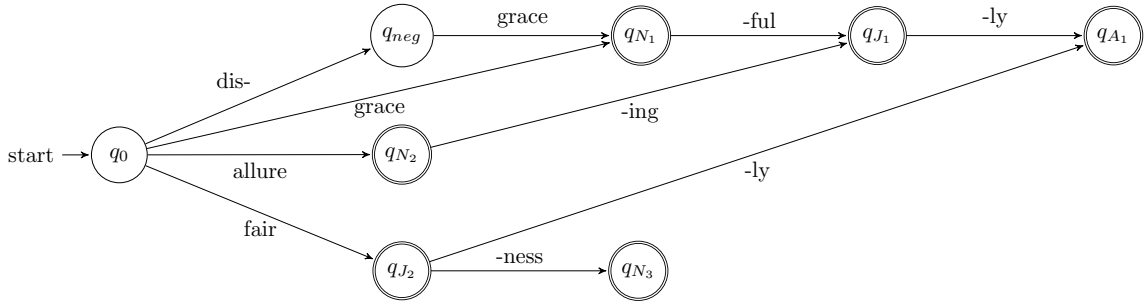



Figure 8.3: Minimization of the finite-state acceptor shown in Figure 8.2.

### 8.1.3 Weighted finite-state acceptors

According to the FSA treatment of morphology, every word is either in or out of the language, with no wiggle room. Perhaps you agree that *musicky* and *fishful* are not valid English words; but if forced to choose, you probably find *a fishful stew* or *a musicky tribute* preferable to *behaving disgracelyful*. Rather than asking whether a word is acceptable, we might like to ask how acceptable it is. Aronoff (1976, page 36) puts it another way: "Though many things are possible in morphology, some are more possible than others." But finite state acceptors gives us no way to express preferences among technically valid choices.

**Weighted finite-state acceptors (WFSAs)** are generalizations of FSAs, in which each accepting path is assigned a score, which is computed from the transitions, the initial state, and the final state. Formally, a weighted finite-state acceptor $M = \langle Q, \Sigma, \lambda, \rho, \delta \rangle$ consists of:

- a finite set of states $Q = \{q_0, q_1, \ldots, q_n\}$;
- a finite alphabet $\Sigma$ of input symbols;
- an initial weight function, $\lambda : Q \rightarrow \mathbb{R}$;
- a final weight function $\rho : Q \rightarrow \mathbb{R}$;
- a transition function $\delta : Q \times \Sigma \times Q \rightarrow \mathbb{R}$.

Thus, WFSAs depart from the FSA formalism in three ways: every state can be an initial state, with score $\lambda(q)$; every state can be an accepting state, with score $\rho(q)$; transitions are possible between any pair of states on any input, with a score $\delta(q_i, \omega, q_j)$. Nonetheless, FSAs can be viewed as a special case: for any FSA $M$ we can build an equivalent WFSA 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 transitions $\{(q_1, \omega) \rightarrow q_2\}$ that are not permitted by the transition function of $M$.

The total score for any path $\pi = t_1, t_2, \ldots, t_N$ is equal to the sum of these scores,

$$d(\pi) = \lambda(\text{from-state}(t_1)) + \sum_n^N \delta(t_n) + \rho(\text{to-state}(t_N)). \quad [8.5]$$

A **shortest-path algorithm** is used to find the minimum-cost path through a WFSA for string $\omega$, with time complexity $\mathcal{O}(E + V \log V)$, where $E$ is the number of edges and $V$ is the number of vertices (Cormen et al., 2009).[2]

#### N-gram language models as WFSAs

§ 5.1 introduced **$n$-gram language models**, in which the probability of a sequence of tokens $w_1, w_2, \ldots, w_M$ is modeled as,

$$\text{p}(w_1, \ldots, w_M) \approx \prod_m^M \text{p}_n(w_m \mid w_{m-1}, \ldots, w_{m-n+1}). \quad [8.6]$$

$N$-gram language models can be modeled as WFSAs. First consider a unigram language model. We need only a single state $q_0$, with transition scores $\delta(q_0, \omega, q_0) = \log \text{p}_1(\omega)$. The initial and final scores can be set to zero. Then the path score for $w_1, w_2, \ldots, w_M$ is equal to,

$$0 + \sum_m^M \delta(q_0, w_m, q_0) + 0 = \sum_m^M \log \text{p}_1(w_m). \quad [8.7]$$

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 $\log \text{p}_2(w_m \mid w_{m-1})$. This means that the transitions scoring function has to somehow "remember" the previous word or words. This can be done by adding more states: to model the bigram probability $\text{p}_2(w_m \mid w_{m-1})$, 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:

$$\begin{aligned}
\delta(q_i, \omega, q_j) &= \begin{cases} \log \Pr_2(w_m = j \mid w_{m-1} = i), & \omega = j \\ \infty, & \omega \neq j \end{cases} \\
\lambda(q_i) &= \log \Pr_2(w_1 = i \mid w_0 = \lozenge) \\
\rho(q_i) &= \log \Pr_2(w_{M+1} = \blacklozenge \mid w_M = i).
\end{aligned}$$

[2]Shortest-path algorithms find the path with the minimum cost. In many cases, the path weights are log probabilities, so we want the path with the maximum score, which can be accomplished by making each local score into a **negative** log-probability. The remainder of this section will refer to **best-path algorithms**, which are assumed to "do the right thing."

The transition function is designed to ensure that the context is recorded accurately: we can move to state $j$ on input $\omega$ only if $\omega = j$; otherwise, transitioning to state $j$ is forbidden by the weight of $-\infty$. The initial weight function $\lambda(q_i)$ is the log probability of receiving $i$ as the first token, and the final weight function $\rho(q_i)$ is the log probability of receiving an "end-of-string" token after observing $w_M = i$. [todo: figure]

#### *Semiring Weighted Finite State Acceptors

The $n$-gram language model WFSA is deterministic: each input has exactly one accepting path, for which the WFSA computes a score. In non-deterministic WFSAs, a given input may have multiple accepting paths, and in some applications, we are interested in the score for the input across all such paths. Such aggregate scores can be computed by generalizing WFSAs with **semiring notation**, first introduced in § 6.4.4.

Let $d(\pi)$ represent the total score for path $\pi = t_1, t_2, \ldots, t_N$, which is computed as,

$$d(\pi) = \lambda(\text{from-state}(t_1)) \otimes \delta(t_1) \otimes \delta(t_2) \otimes \ldots \otimes \delta(t_N) \otimes \rho(\text{to-state}(t_N)). \quad [8.8]$$

This is a generalization of Equation 8.5 to semiring notation, using the semiring multiplication operator $\otimes$ in place of addition.

Now let $s(\omega)$ represent the total score for all paths $\Pi(\omega)$ that consume input $\omega$,

$$s(\omega) = \bigoplus_{\pi \in \Pi(\omega)} d(\pi). \quad [8.9]$$

Here, semiring addition ($\oplus$) is used to combine the scores of multiple paths.

The generalization to semirings covers a number of useful special cases. In the log-probability semiring, multiplication is defined as $\log \text{p}(x) \otimes \log \text{p}(y) = \log \text{p}(x) + \log \text{p}(y)$, and addition is defined as $\log \text{p}(x) \oplus \log \text{p}(y) = \log(\text{p}(x) + \text{p}(y))$. Thus, $s(\omega)$ represents the log-probability of accepting input $\omega$, marginalizing over all paths $\pi \in \Pi(\omega)$. In the **boolean semiring**, the $\otimes$ operator is logical conjunction, and the $\oplus$ operator is logical disjunction. This reduces to the special case of unweighted finite-state acceptors, where the score $s(\omega)$ is a boolean indicating whether there exists any accepting path for $\omega$. In the **tropical semiring**, the $\oplus$ operator is a maximum, so the resulting score is the score of the best-scoring path through the WFSA. The `OpenFST` toolkit uses semirings and polymorphism to implement general algorithms for weighted finite-state automata (Allauzen et al., 2007).

#### *Interpolated $n$-gram language models

Recall from § 5.2.3 that an **interpolated $n$-gram language model** combines the probabilities from multiple $n$-gram models. For example, an interpolated bigram language model

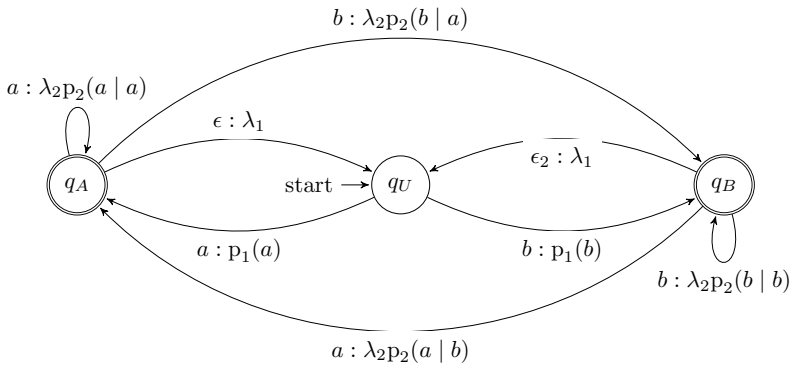



Figure 8.4: WFSA implementing an interpolated bigram/unigram language model, on the alphabet $\Sigma = \{a, b\}$. For simplicity, the WFSA is contrained to force the first token to be generated from the unigram model, and does not model the emission of the end-of-sequence token.

computes probability,

$$\hat{\mathrm{p}}(w_m \mid w_{m-1}) = \lambda_1 \mathrm{p}_1(w_m) + \lambda_2 \mathrm{p}_2(w_m \mid w_{m-1}), \tag{8.10}$$

with $\hat{\mathrm{p}}$ indicating the interpolated probability, $\mathrm{p}_2$ indicating the bigram probability, and $\mathrm{p}_1$ indicating the unigram probability. We set $\lambda_2 = (1 - \lambda_1)$ so that the probabilities sum to one.

Knight and May (2009) describe how to implement an interpolated bigram language model using a non-deterministic WFSA. The basic idea is shown in Figure 8.4. In the WFSA implementation of the bigram language model, we have one state for each element in the vocabulary — in this case, the states $q_A$ and $q_B$ — which are capture the contextual conditioning in the bigram probabilities. To model unigram probabilities, we add an additional state $q_U$, which "forgets" the context. Transitions out of $q_U$ involve unigram probabilities, $\mathrm{p}_1(a)$ and $\mathrm{p}_2(b)$; transitions into $q_U$ emit the empty symbol $\epsilon$, and have probability $\lambda_1$, reflecting the interpolation weight for the unigram model. The interpolation weight for the bigram model is included directly in the weight of the transition $q_A \rightarrow q_B$.

The epsilon transitions into $q_U$ make this WFSA non-deterministic. Consider the score for the sequence $(a, b, b)$. The initial state is $q_U$, so the symbol $a$ is generated with score $\mathrm{p}_1(a)$[3] Next, we can generate $b$ from the unigram model by taking the transition $q_A \rightarrow q_B$, with score $\lambda_2 \mathrm{p}_2(b \mid a)$. Alternatively, we can take a transition back to $q_U$ with score $\lambda_1$, and then emit $b$ from the unigram model with score $\mathrm{p}_1(b)$. To generate the final $b$ token,

[3] We could model the sequence-initial bigram probability $\mathrm{p}_2(a \mid \square)$, but for simplicity the WFSA does not admit this possibility, which would require another state.

we face the same choice: emit it directly from the self-transition to $q_B$, or transition to $q_U$ first.

The total score for the sequence $(a, b, b)$ is the semiring sum over all accepting paths,

$$\begin{aligned} s(a,b,b) = &\big(\mathrm{p}_1(a) \otimes \lambda_2 \mathrm{p}_2(b \mid a) \otimes \lambda_2 \mathrm{p}(b \mid b)\big) \\ &\oplus \big(\mathrm{p}_1(a) \otimes \lambda_1 \otimes \mathrm{p}_1(b) \otimes \lambda_2 \mathrm{p}(b \mid b)\big) \\ &\oplus \big(\mathrm{p}_1(a) \otimes \lambda_2 \mathrm{p}_2(b \mid a) \otimes \mathrm{p}_1(b) \otimes \mathrm{p}_1(b)\big) \\ &\oplus \big(\mathrm{p}_1(a) \otimes \lambda_1 \otimes \mathrm{p}_1(b) \otimes \mathrm{p}_1(b) \otimes \mathrm{p}_1(b)\big) . \end{aligned} \quad [8.11]$$

Each line in Equation 8.11 represents the probability of a specific path through the WFSA. In the probability semiring, $\otimes$ is multiplication, so that each path is the product of each transition weight, which are themselves probabilities. The $\oplus$ operator is addition, so that the total score is the sum of the scores (probabilities) for each path. This corresponds to the probability under the interpolated bigram language model.

### 8.1.4 Finite-state transducers

Finite state acceptors can determine whether a string is in a regular language, and weighted finite state acceptors can compute a score for every string over a given alphabet. **Finite-state transducers** (FSTs) extend the formalism further, by adding an output symbol to each transition. Formally, we write $T = \langle Q, \Sigma, \Omega, \lambda, \rho, \delta \rangle$, with $\Omega$ representing an output vocabulary and the transition function $\delta : Q \times (\Sigma \cup \epsilon) \times (\Omega \cup \epsilon) \times Q \rightarrow \mathbb{R}$ mapping from states, input symbols, and output symbols to states. The remaining elements $(Q, \Sigma, \lambda, \rho)$ are identical to their definition in weighted finite state acceptors (§ 8.1.3). Thus, each path through the FST $T$ transduces the input string into an output.

#### String edit distance

The **edit distance** between two strings $s$ and $t$ is a measure of how many basic 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,

$$\delta(q, a, a, q) = \delta(q, b, b, q) = 0 \quad [8.12]$$
$$\delta(q, a, b, q) = \delta(q, b, a, q) = 1 \quad [8.13]$$
$$\delta(q, a, \epsilon, q) = \delta(q, b, \epsilon, q) = 1 \quad [8.14]$$
$$\delta(q, \epsilon, a, q) = \delta(q, \epsilon, b, q) = 1. \quad [8.15]$$

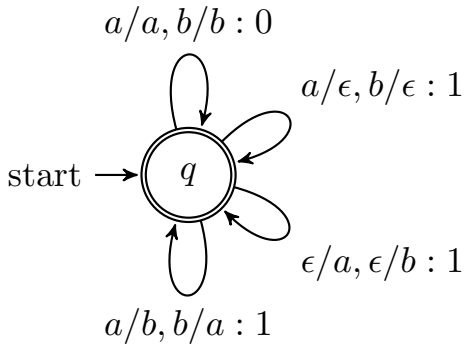



Figure 8.5: State diagram for the Levenshtein edit distance finite state transducer. The label $x/y : c$ indicates a cost of $c$ for a transition with input $x$ and output $y$.

The state diagram is shown in Figure 8.5.

For a given string pair, there are multiple paths through the transducer: the best-scoring path from *dessert* to *desert* involves a single deletion, for a total score of 1; the worst-scoring path has a score of 13. In edit distance, we are interested in the lowest score, but as we will see, in other applications, the score is computed by marginalizing over paths.

#### The Porter stemmer

The Porter (1980) stemming algorithm was discussed in chapter 3. It is a "lexicon-free" algorithm for stripping suffixes from English words, using a sequence of character-level rules. Each rule can be described by an unweighted finite-state transducer. The first rule is:

$$\textit{-sses} \rightarrow \textit{-ss} \quad \text{e.g., } \textit{dresses} \rightarrow \textit{dress} \tag{8.16}$$
$$\textit{-ies} \rightarrow \textit{-i} \quad \text{e.g., } \textit{parties} \rightarrow \textit{parti} \tag{8.17}$$
$$\textit{-ss} \rightarrow \textit{-ss} \quad \text{e.g., } \textit{dress} \rightarrow \textit{dress} \tag{8.18}$$
$$\textit{-s} \rightarrow \epsilon \quad \text{e.g., } \textit{cats} \rightarrow \textit{cat} \tag{8.19}$$

The final two lines appear to conflict; they are meant to be interpreted as an instruction to remove a terminal *-s* unless it is part of an *-ss* ending. A state diagram to handle just these final two lines is shown in Figure 8.6. Be sure to understand how this finite-state transducer handles *cats*, *steps*, *bass*, and *basses*.

#### Inflectional morphology

In **inflectional morphology**, word **lemmas** are modified to add grammatical information such as tense, number, and case. For example, many English nouns are pluralized by the -

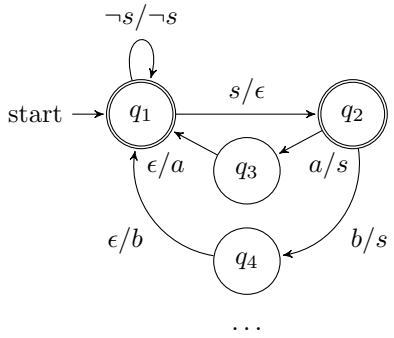



Figure 8.6: State diagram for final two lines of step 1a of the Porter stemming diagram. States $q_3$ and $q_4$ "remember" the observations $a$ and $b$ respectively; the ellipsis ... represents additional states for each symbol in the input alphabet. The shorthand notation $\neg s/\neg s$ is not part of the FST formalism, but should be interpreted as 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 | vivís |
| ellos, ellas (3rd plural); ustedes (2nd plural) | cantan | comen | viven |

Table 8.1: 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.

*s* suffix, and many verbs are converted to past tense by the *-ed* suffix. English's inflectional morphology is considerably simpler than many of the world's languages. For example, Romance languages (derived from Latin) feature complex systems of verb suffixes which must agree with the person and number of the verb, as shown in Table 8.1.

The task of **morphological analysis** is to read a form like *canto*, and output an analysis like CANTAR+VERB+PRESIND+1P+SING, where +PRESIND describes the tense as present indicative, +1P indicates the first-person, and +SING indicates the singular number. The task of **morphological generation** is the reverse, going from CANTAR+VERB+PRESIND+1P+SING to *canto*. Finite-state transducers are an attractive solution, because they can solve both problems with a single model (Beesley and Karttunen, 2003). As an example, Figure 8.7

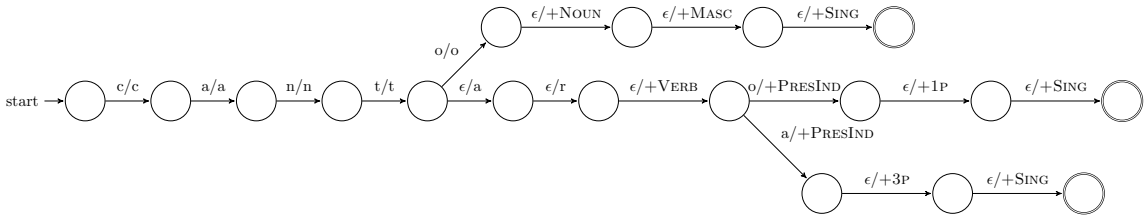



Figure 8.7: Fragment of a finite-state transducer for Spanish morphology. There are two accepting paths for the input *canto*: *canto*+NOUN+MASC+SING (masculine singular noun, meaning a song), and *cantar*+VERB+PRESIND+1P+SING (I sing). There is also an accepting path for *canta*, with output *cantar*+VERB+PRESIND+3P+SING (he/she sings).

shows a fragment of a finite-state transducer for Spanish inflectional morphology. The input vocabulary $\Sigma$ corresponds to the set of letters used in Spanish spelling, and the output vocabulary $\Omega$ corresponds to these same letters, plus the vocabulary of morphological features (e.g., +SING, +VERB). In Figure 8.7, there are two paths that take *canto* as input, corresponding to the verb and noun meanings; the choice between these paths could be guided by a part-of-speech tagger. By **inversion**, the inputs and outputs for each transition are switched, resulting in a finite-state generator, capable of producing the correct **surface form** for any morphological analysis.

Finite-state morphological analyzers and other unweighted transducers can be designed by hand. The designer's goal is to avoid **overgeneration** — accepting strings or making transductions that are not valid in the language — as well as **undergeneration** — failing to accept strings or transductions that are valid. For example, a pluralization transducer that does not accept *foot/feet* would undergenerate. Suppose we "fix" the transducer to accept this example, but as a side effect, it now accepts *boot/beet*; the transducer would then be said to overgenerate. A transducer that accepts *foot/foots* but not *foot/feet* would both overgenerate and undergenerate.

### Finite-state composition

Designing finite-state transducers to capture the full range of morphological phenomena in any real language is a huge task. In computer science, **modularization** is a key technique for handling such issues — decompose a large and unwieldly problem into a set of subproblems, each of which will hopefully have a concise solution. In finite-state automata, this can be performed by by **composition**: feeding the output of one transducer $T_1$ as the input to another transducer $T_2$, written $T_2 \circ T_1$. Formally, if there exists some $y$ such that $(x, y) \in T_1$ (meaning that $T_1$ produces output $y$ on input $x$), and $(y, z) \in T_2$, then $(x, z) \in (T_2 \circ T_1)$. Because finite-state transducers are **closed** under composition, there is guaranteed to be a single finite-state transducer that $T_3 = T_2 \circ T_1$, which can be constructed as a machine with one state for each pair of states in $T_1$ and $T_2$ (Mohri et al.,

2002).

**Example: Morphology and orthography** In English morphology, the suffix *-ed* is added to signal the past tense for many verbs: *cook→cooked*, *want→wanted*, etc. However, English **orthography** dictates that this process cannot produce a spelling with consecutive *e*'s, so that *bake→baked*, not *bakeed*. A modular solution is to build separate transducers for morphology and orthography. The morphological transducer $T_M$ transduces from *bake*+PAST to *bake+ed*, with the + symbol indicating a segment boundary. The input alphabet of $T_M$ includes the lexicon of words and the set of morphological features; the output alphabet includes the characters *a*-*z* and the + boundary marker. Next, an orthographic transducer $T_O$ is responsible for the transductions *cook+ed* $\rightarrow$ *cooked*, and *bake+ed* $\rightarrow$ *baked*. The input alphabet of $T_O$ must be the same as the output alphabet for $T_M$, and the output alphabet is simply the characters *a*-*z*. The composition ($T_O \circ T_M$ then transduces from *bake*+PAST to the spelling *baked*. The design of $T_O$ is left as an exercise.

**Example: Hidden Markov models** Hidden Markov models (chapter 6) can be viewed as weighted finite-state transducers. Recall that a hidden Markov model defines a joint probability over words and tags, $\text{p}(\boldsymbol{w}, \boldsymbol{y})$, 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,

$$\begin{aligned}\delta(q_{m-1,i}, Y_m = j, q_{m,j}) &= \log \text{p}(w_m, Y_m = j \mid Y_{m-i} = j) && [8.20]\\ &= \log \text{p}(w_m \mid Y_m = j) + \log \Pr(Y_m = j \mid Y_{m-1} = i). && [8.21]\end{aligned}$$

Because there is only one possible transition for each tag $Y_m$, this WFSA is deterministic. The score for any tag sequence $\{y_m\}_{m=1}^M$ is the sum of these log-probabilities, corresponding to the total log probability $\log \text{p}(\boldsymbol{w}, \boldsymbol{y})$.

Furthermore, the trellis can be constructed by the composition of simpler FSTs.

- First, take construct a "transition" transducer to represent a bigram probability model over tag sequences, $T_T$. This transducer is almost identical to the $n$-gram language model acceptor in § 8.1.3: there is one state for each tag, and the edge weights equal to the transition log-probabilities, $\delta(q_i, j, j, q_j) = \log \Pr(Y_m = j \mid Y_{m-1} = i)$. Note that $T_T$ is a transducer, with identical input and output at each arc; this makes it possible to compose $T_T$ with other transducers.
- Next, construct an "emission" transducer to represent the probability of words given tags, $T_E$. This transducer has only a single state, with arcs for each word/tag pair, $\delta(q_0, i, j, q_0) = \log \Pr(W_m = j \mid Y_m = i)$. The input vocabulary is the set of all tags, and the output vocabulary is the set of all words.

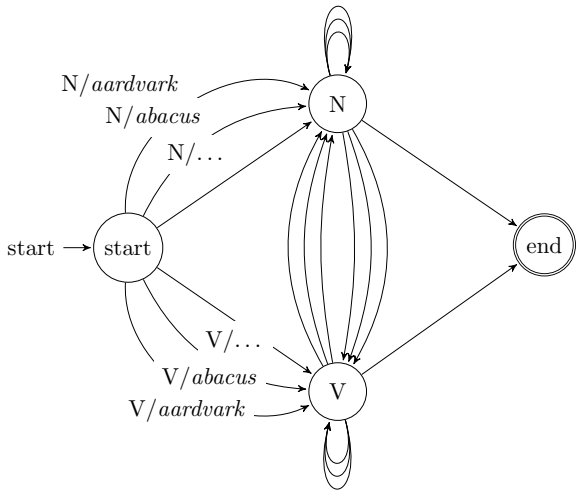



Figure 8.8: Finite-state transducer for hidden Markov models, with a small tagset of **n**ouns and **v**erbs. For each pair of tags (including self-loops), there is an edge for every word in the vocabulary. For simplicity, input and output are only shown for the edges from the start state. Weights are also omitted from the diagram; for each edge from $q_i$ to $q_j$, the weight is equal to $\log \text{p}(w_m, Y_m = j \mid Y_{m-1} = i)$, except for edges to the end state, which are equal to $\log \Pr(Y_m = \blacklozenge \mid Y_{m-1} = i)$.

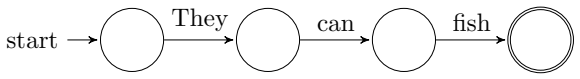



Figure 8.9: Chain finite-state acceptor for the input *They can fish*.

- The composition $T_E \circ T_T$ is a finite-state transducer with one state per tag, as shown in Figure 8.8. Each state has $V \times K$ outgoing edges, representing transitions to each of the $K$ other states, with outputs for each of the $V$ words in the vocabulary. The weights for these edges are equal to $\delta(q_i, Y_m = j, w_m, q_j) = \log \text{p}(w_m, Y_m = j \mid Y_{m-1} = i)$.

- Recall the **trellis** from § 6.3, a structure with $M \times K$ nodes, for each of the $M$ words to be tagged and each of the $K$ tags in the tagset. The trellis can be built by composition of $(T_E \circ T_T)$ against an unweighted **chain FSA** $M_A(\boldsymbol{w})$ that is specially constructed to accept only a given input $w_1, w_2, \ldots, w_M$, shown in Figure 8.9. The trellis for input $\boldsymbol{w}$ is built from the composition $M_A(\boldsymbol{w}) \circ (T_E \circ T_T)$. Composing with the unweighted $M_A(\boldsymbol{w})$ does not affect the edge weights from $(T_E \circ T_T)$, but it selects the subset of paths that generate the word sequence $\boldsymbol{w}$.

### 8.1.5 *Learning weighted finite-state automata

In generative models such as $n$-gram language models and hidden Markov models, the edge weights correspond to log probabilities that can be obtained from relative frequency estimation. However, in other cases, we wish to learn the edge weights from input/output pairs. This is difficult in non-deterministic finite-state automata, because we do not observe the specific arcs that are traversed in accepting the input, or in transducing from input to output. The path through the automaton is a **latent variable**.

chapter 4 presented one solution for learning with latent variables: expectation maximization (EM). This involves computing a distribution $q(\cdot)$ over the latent variable, and iterating between updates to this distribution and updates to the parameters — in this case, the arc weights. The **forward-backward algorithm** (§ 6.5.3) describes a dynamic 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 **hidden variable perceptron**. The model is described formally in § 11.4, 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.

## 8.2 Context-free languages

Beyond the restricted class of regular languages lie the context-free languages. An example of a language that is context-free but not finite-state is the language 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]

[4]Details of the proof can be found in introductory computer science theory textbooks (e.g., Sipser, 2012).

| | | | | |
|---|---|---|---|---|
| | | the dog | | |
| | the cat | the dog | chased | |
| the goat | the cat | the dog | chased | kissed |
| | | ... | | |

Figure 8.10: Three levels of center embedding

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 8.10. 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. English includes center embedding, and so the argument goes, English grammar as a whole cannot be regular.[5]

A more practical argument for moving beyond regular languages is **modularity**. Many linguistic phenomena — especially in syntax — involve constraints that apply at long distance. Consider the problem of determiner-noun number agreement in English: we can say *the coffee* and *these coffees*, but not **these coffee*. By itself, this is easy enough to model in an FSA. However, fairly complex modifying expressions can be inserted between the determiner and the noun:

(8.5) the burnt coffee

(8.6) the badly-ground coffee

(8.7) the burnt and badly-ground Italian coffee

(8.8) these burnt and badly-ground Italian coffees

(8.9) *these burnt and badly-ground Italian coffee

[5]The 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.

Again, an FSA can be designed to accept modifying expressions such as *burnt and badly-ground Italian*. Let's call this FSA $F_M$. To reject the final example, a finite-state acceptor must somehow "remember" that the determiner was plural when it reaches the noun *coffee* at the end of the expression. The only way to do this is to make two identical copies of $F_M$: one for singular determiners, and one for plurals. While this is possible in the finite-state framework, it is inconvenient — especially in languages where more than one attribute of the noun is marked by the determiner. **Context-free languages** are a superset of regular languages, which include cases such as $a^n b^n$, and which enable modularity even in the case of long-range dependencies such as determiner-noun agreement.

### 8.2.1 Context-free grammars

Context-free languages are specified by **context-free grammars (CFGs)**, which are tuples $\langle N, \Sigma, R, S \rangle$ consisting of:

- a finite set of **non-terminals** $N$;
- a finite alphabet $\Sigma$ of **terminal symbols**;
- a set of **production rules** $R$, each of the form $A \rightarrow \beta$, where $A \in N$ and $\beta \in (\Sigma \cup N)^*$;
- a designated start symbol $S$.

In the production rule $A \rightarrow \beta$, the left-hand side (LHS) $A$ must be a non-terminal; the right-hand side (RHS) can be a sequence of terminals or non-terminals, $\{n, \sigma\}^*, n \in N, \sigma \in \Sigma$. A given non-terminal can appear on the left-hand side of many production rules. The non-terminal $A$ on the left-hand side can also appear on the right-hand side; such recursive productions are analogous to self-loops in finite-state automata. Context-free grammars are "context-free" in the sense that the production rule depends only on the LHS, and not on its ancestors or neighbors; this is analogous to Markov property of finite-state automata, in which the behavior at each step depends only on the current state, on not on the path by which that state was reached.

A **derivation** $\tau$ is a sequence of steps from the start symbol $S$ to a surface string $\boldsymbol{w} \in \Sigma^*$, which is the **yield** of the derivation. A string $\boldsymbol{w}$ is in a context-free language if there is some derivation from $S$ yielding $\boldsymbol{w}$. **Parsing** is the problem of finding a derivation for a string in a grammar. Algorithms for parsing are described in chapter 9.

Context-free grammars are analogous to regular expressions, in that they define the language but not the computation necessary to recognize it. The context-free analogues to finite-state acceptors are **pushdown automata**, a theoretical model of computation in which input symbols can be pushed onto a stack with potentially infinite depth (Sipser, 2012).

$$
\begin{aligned}
\text{S} &\rightarrow \text{S OP S} \mid \text{NUM} \\
\text{OP} &\rightarrow + \mid - \mid \times \mid \div \\
\text{NUM} &\rightarrow \text{NUM DIGIT} \mid \text{DIGIT} \\
\text{DIGIT} &\rightarrow \texttt{0} \mid \texttt{1} \mid \texttt{2} \mid \ldots \mid \texttt{9}
\end{aligned}
$$

Figure 8.11: A context free grammar for arithmetic expressions

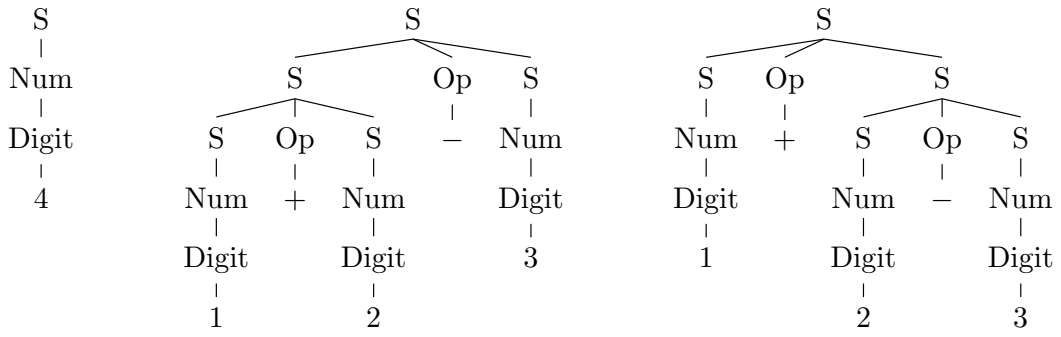



Figure 8.12: Some example derivations from the arithmetic grammar in Figure 8.11

**Example**

Figure 8.11 shows a context-free grammar for arithmetic expressions such as $(1 + 2) \div (3 \times 4)$. In this grammar, the terminal symbols include the digits $\{\texttt{1}, \texttt{2}, \ldots, \texttt{9}\}$ and the operators $\{+, -, \times, \div\}$. The rules include the $\mid$ symbol, a notational convenience that makes it possible to specify multiple right-hand sides on a single line: the statement $A \rightarrow x \mid y$ defines *two* productions, $A \rightarrow x$ and $A \rightarrow y$. This grammar is recursive: the non-termals S and NUM can produce themselves.

Derivations are typically shown as trees, with production rules applied from the top to the bottom. The tree on the left in Figure 8.12 describes the derivation of a single digit, through the sequence of productions S $\rightarrow$ EXPR $\rightarrow$ NUM $\rightarrow$ DIGIT $\rightarrow$ 4 (these are all **unary productions**, because the right-hand side contains a single element). The other two trees in Figure 8.12 show alternative derivations of the string $\texttt{1} + \texttt{2} - \texttt{3}$. The existence of multiple derivations for a string indicates that the grammar is **ambiguous**.

Context-free derivations can also be written out according to the pre-order tree traver-

sal. For the two derivations of `1 + 2 - 3` in Figure 8.12, the notation is:

(S (S (S (Num (Digit 1))) (Op +) (S (Num (Digit 2)))) (Op - ) (S (Num (Digit 3)))) [8.22]

(S (S (Num (Digit 1))) (Op +) (S (Num (Digit 2)) (Op - ) (S (Num (Digit 3))))). [8.23]

**Grammar equivalence and Chomsky Normal Form**

A single context-free language may be expressed by more than one context-free grammar: as a trivial example, the language $a^n b^n$ ($n > 0$) can be recognized by either of the following grammars:

$$S \rightarrow aSb \mid ab$$
$$S \rightarrow aSb \mid aabb \mid ab$$

Two grammars are **weakly equivalent** if they generate the same strings. Two grammars are **strongly equivalent** if they generate the same strings via the same derivations. (The grammars above are only weakly equivalent.)

In **Chomsky Normal Form (CNF)**, the right-hand side of every production includes either two non-terminals, or a single terminal symbol:

$$A \rightarrow BC$$
$$A \rightarrow a$$

All CFGs can be converted into a CNF grammar that is weakly equivalent. To convert a grammar into CNF, we first address productions that have more than two non-terminals on the RHS by creating new "dummy" non-terminals. For example, if we have the production,

$$\text{W} \rightarrow \text{X Y Z}, \quad [8.24]$$

we can replace it with two productions,

$$\text{W} \rightarrow \text{X W}\backslash\text{X} \quad [8.25]$$
$$\text{W}\backslash\text{X} \rightarrow \text{Y Z}. \quad [8.26]$$

In these productions, W\X is a new dummy non-terminal, indicating a phrase that would be W if its left neighbor is an X. This transformation **binarizes** the grammar, which is critical for efficient bottom-up parsing (chapter 9). Productions whose right-hand side contains a mix of terminal and non-terminal symbols can be replaced in a similar fashion.

Unary non-terminal productions $A \rightarrow B$ are replaced as follows: identify all productions $B \rightarrow \alpha$, and add $A \rightarrow \alpha$ to the grammar. For example, in the grammar described in Figure 8.11, we would replace NUM $\rightarrow$ DIGIT with NUM $\rightarrow$ `1` | `2` | ... | `9`. However, we keep the production NUM $\rightarrow$ NUM DIGIT, which is a valid binary production.

### 8.2.2 Natural language syntax as a context-free language

Context-free grammars are widely used to represent **syntax**, which is the system of constraints and preferences that determine whether an utterance is judged to be grammatical. If this representation were perfectly faithful, then a natural language such as English could be transformed into a formal language, consisting of exactly the (infinite) set of strings that would be judged to be grammatical by a fluent English speaker. We could then build parsing software that would automatically determine if a given utterance were grammatical.[6]

Contemporary theories generally do *not* consider natural languages to be context-free, yet context-free grammars are widely used in automated natural language parsing. The reason is that context-free representations strike a good balance: it is possible to represent a broad range of syntactic phenomena in context-free form, and it is possible to parse context-free languages efficiently. This section therefore describes how to handle a core fragment of English syntax in context-free form, following the conventions of the **Penn Treebank** (PTB; Marcus et al., 1993), a large-scale annotation of English language syntax. The generalization to "mildly" context sensitive languages is discussed in § 8.3.

The Penn Treebank annotation is a **phrase-structure grammar** of English. This means that sentences are broken down into **constituents**, which are contiguous sequences of words that function as coherent units for the purpose of linguistic analysis. Constituents can be:

- moved around the sentence

  (8.10) Abigail gave (her brother) (a fish).

  (8.11) Abigail gave (a fish) to (her brother).

- substituted

  (8.12) Max thanked (his older sister).

  (8.13) Max thanked (her).

- coordinated

  (8.14) (Abigail) and (her younger brother) bought a fish.

  (8.15) Abigail (bought a fish) and (gave it to Max).

  (8.16) Abigail (bought) and (greedily ate) a fish.

[6] As with morphology, it is impossible to do justice to natural language syntax here; the reader is encouraged to consult a text on linguistics (Akmajian et al., 2010; Bender, 2013).

These examples argue for units such as *her brother* and *bought a fish* to be treated as constituents. Other sequences of words in these examples, such as *Abigail gave* and *brother a fish*, cannot be moved, substituted, and coordinated in these ways. In phrase-structure grammar, constituents are nested, so that *the senator from New Jersey* contains the constituent *from New Jersey*, which in turn contains *New Jersey*. The sentence itself is the maximal constituent; each word is a minimal constituent, derived from a unary production from a part-of-speech tag. Between part-of-speech tags and sentences are **phrases**. In phrase-structure grammar, phrases are typically defined as the **projection** of a **head word**: for example, a **noun phrase** corresponds to a noun and the group of words that modify it, such as *her younger* ***brother***; a **verb phrase** includes the verb and its modifiers, such as ***bought*** *a fish* and *greedily* ***ate*** *it*.

In context-free grammars, constituents correspond to the non-terminals. Grammar design involves choosing the right set of non-terminals. Fine-grained non-terminals make it possible to represent more fine-grained linguistic phenomena. For example, by distinguishing singular and plural noun phrases, it is possible to design a grammar rule that enforces subject-verb agreement in English. However, enforcing subject-verb agreement is considerably more complicated in languages like Spanish, where the verb must agree in both person and number with subject. In general, grammar designers must trade off between **overgeneration** — a grammar that permits ungrammatical sentences — and **undergeneration** — a grammar that fails to permit grammatical sentences. Furthermore, if the grammar is to support manual annotation of syntactic structure, it must be simple enough to annotate efficiently.

### 8.2.3 A phrase-structure grammar for English

To better understand how phrase-structure grammar works, let's consider the specific case of the Penn Treebank grammar of English. The main phrase categories in the Penn Treebank (PTB) are based on the main part-of-speech classes: noun phrase (NP), verb phrase (VP), prepositional phrase (PP), adjectival phrase (ADJP), and adverbial phrase (ADVP). The top-level category is S, which conveniently stands in for both "sentence" and the "start" symbol. **Complement clauses** (e.g., *She said* ***that she liked spicy food***) are represented by the non-terminal SBAR. The terminal symbols in the grammar are individual words, which come from unary productions from part-of-speech tags (the PTB tagset is described in § 7.1).

This section explores the productions from the major phrase-level categories, explaining how to generate individual tag sequences. The production rules are approached in a "theory-driven" manner: first the syntactic properties of each phrase type are described, and then some of the necessary production rules are listed. But it is important to keep in mind that the Penn Treebank was produced in a "data-driven" manner. After the set of non-terminals was specified, annotators were free to analyze each sentence in what-

ever way seemed most linguistically accurate, subject to some high-level guidelines. The grammar of the Penn Treebank is simply the set of productions that were required to analyze the several million words of the corpus. By design, the grammar does not exclude ungrammatical sentences.

**Sentences**

The most common production rule for sentences is,

$$\text{S} \rightarrow \text{NP VP} \quad [8.27]$$

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:

$$\text{S} \rightarrow \text{ADVP NP VP} \quad \textit{Unfortunately Abigail ate the kimchi.} \quad [8.28]$$
$$\text{S} \rightarrow \text{S CC S} \quad \textit{Abigail ate the kimchi and Max had a burger.} \quad [8.29]$$
$$\text{S} \rightarrow \text{VP} \quad \textit{Eat the kimchi.} \quad [8.30]$$

where ADVP is an adverbial phrase (e.g., *unfortunately*, *very unfortunately*) and CC is a coordinating conjunction (e.g., *and*, *but*).[7]

**Noun phrases**

Noun phrases refer to entities, real or imaginary, physical or abstract: *Asha*, *the steamed dumpling*, *parts and labor*, *nobody*, 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:

$$\text{NP} \rightarrow \text{NN} \mid \text{NNS} \mid \text{NNP} \mid \text{PRP} \quad [8.31]$$
$$\text{NP} \rightarrow \text{DET NN} \mid \text{DET NNS} \mid \text{DET NNP} \quad [8.32]$$

The part-of-speech tags NN, NNS, and NNP refer to singular, plural, and proper nouns; PRP refers to personal pronouns, and DET refers to determiners. The grammar also contains terminal productions from each of these tags, e.g., PRP $\rightarrow$ *I* | *you* | *we* | ....

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:

$$\text{NP} \rightarrow \text{ADJP NN} \mid \text{ADJP NNS} \mid \text{DET ADJP NN} \mid \text{DET ADJP NNS} \quad [8.33]$$
$$\text{NP} \rightarrow \text{CD NNS} \mid \text{DET CD NNS} \mid \ldots \quad [8.34]$$

[7] Notice 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.

Some noun phrases include multiple nouns, such as *the liberation movement* and *an antelope horn*, necessitating additional productions:

$$\text{NP} \rightarrow \text{NN NN} \mid \text{NN NNS} \mid \text{DET NN NN} \mid \ldots \quad [8.35]$$

These multiple noun constructions can be combined with adjectival phrases and cardinal 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 → NP CC NP | *e.g., the red and the black* | [8.36] |
| NP → NP PP | *e.g., the President of the Georgia Institute of Technology* | [8.37] |
| NP → NP SBAR | *e.g., the bicycle that I found outside* | [8.38] |
| NP → NP VP | *e.g., a bicycle made of titanium* | [8.39] |

These recursive productions are a major source of ambiguity, because the VP and PP nonterminals can also generate NP children. Thus, the *the President of the Georgia Institute of Technology* can be derived in two ways, as can *a bicycle made of titanium found outside*.

But aside from these few recursive productions, the noun phrase fragment of the Penn Treebank grammar is relatively flat, containing a large of number of productions that go from NP directly to a sequence of parts-of-speech. If noun phrases had more internal structure, the grammar would need fewer rules; as we will see, this would make parsing faster and machine learning easier. Vadas and Curran (2011) propose to add additional structure in the form of a new non-terminal called a **nominal modifier** (NML), e.g.,

(8.17) `(NP (NN crude) (NN oil) (NNS prices))` (PTB analysis)

(8.18) `(NP (NML (NN crude) (NN oil)) (NNS prices))` (NML-style analysis)

Another proposal is to treat the determiner as the head of a **determiner phrase** (DP; Abney, 1987). There are linguistic arguments for and against determiner phrases (Van Eynde, 2006, e.g.,). From the perspective of context-free grammar, DPs enable more structured analyses of some constituents, e.g.,

(8.19) `(NP (DT the) (JJ bald) (NN man))` (PTB analysis)

(8.20) `(DP (DT the) (NP (JJ bald) (NN man)))` (DP-style analysis)

#### 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 (VBZ; *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*).[8] Each of these forms can constitute a verb phrase on its own:

$$\text{VP} \rightarrow \text{VB} \mid \text{VBZ} \mid \text{VBD} \mid \text{VBN} \mid \text{VBG} \mid \text{VBP} \quad [8.40]$$

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 infitival *to* (TO):

| | | |
|---|---|---|
| VP → MD VP | *She* ***will snack*** | [8.41] |
| VP → VBD VP | *She* ***had snacked*** | [8.42] |
| VP → VBZ VP | *She* ***has been snacking*** | [8.43] |
| VP → VBN VP | *She has* ***been snacking*** | [8.44] |
| VP → TO VP | *She wants* ***to snack*** | [8.45] |
| VP → VP VP | *She* ***buys and eats*** *many snacks* | [8.46] |

Each of these productions uses recursion, with VP appearing on the right-hand side. This enables the creation of very complex verb phrases, such as *She will have wanted to have been snacking*.

Transitive verbs take noun phrases as direct objects, and ditransitive verbs take two direct objects:

| | | |
|---|---|---|
| VP → VBZ NP | *She* ***teaches algebra*** | [8.47] |
| VP → VBG NP | *She has been* ***teaching algebra*** | [8.48] |
| VP → VBD NP NP | *She taught her brother algebra* | [8.49] |

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:

| | | |
|---|---|---|
| VP → VBZ S | *Asha* ***wants to eat the kimchi*** | [8.50] |
| VP → VBZ SBAR | *Asha* ***knows that Boyang eats the kimchi*** | [8.51] |

The first production overgenerates, licensing sentences like **Asha sees Boyang eats the kimchi*. This problem could be addressed by designing a more specific set of sentence nonterminals, indicating whether the main verb can be conjugated.

[8] It bears emphasis the principles governing this tagset design are entirely English-specific: VBP is a meaningful category only because English morphology distinguishes third-person singular from all person-number combinations.

Verbs can also be modified by prepositional phrases and adverbial phrases:

| | | |
|---|---|---|
| VP → VBZ PP | *She* ***studies at night*** | [8.52] |
| VP → VBZ ADVP | *She* ***studies intensively*** | [8.53] |
| VP → ADVP VBG | *She is* ***not studying*** | [8.54] |

Again, because these productions are not recursive, the grammar must include productions for every verb part-of-speech.

A special set of verbs, known as **copula**, can take **predicative adjectives** as direct objects:

| | | |
|---|---|---|
| VP → VBZ ADJP | *She* ***is hungry*** | [8.55] |
| VP → VBP ADJP | *Success* ***seems increasingly unlikely*** | [8.56] |

The PTB does not have a special non-terminal for copular verbs, so this production generates 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 → VB PRT | *She told them to* ***fuck off*** | [8.57] |
| VP → VBD PRT NP | *They* ***gave up their ill-gotten gains*** | [8.58] |

As the second production shows, particle productions are required for all configurations of verb parts-of-speech and direct objects.

#### Other contituents

The remaining constituents require far fewer productions. **Prepositional phrases** almost always consist of a preposition and a noun phrase,

| | | |
|---|---|---|
| PP → IN NP | *United States* ***of America*** | [8.59] |
| PP → TO NP | *He gave his kimchi* ***to Abigail*** | [8.60] |

Similarly, complement clauses consist of a complementizer (usually a preposition, possibly null) and a sentence,

| | | |
|---|---|---|
| SBAR → IN S | *She said* ***that it was spicy*** | [8.61] |
| SBAR → S | *She said* ***it was spicy*** | [8.62] |

Adverbial phrases are usually bare adverbs (ADVP → RB), with a few exceptions:

| | | |
|---|---|---|
| ADVP → RB RBR | *They went* ***considerably further*** | [8.63] |
| ADVP → ADVP PP | *They went* ***considerably further than before*** | [8.64] |

The tag RBR is a comparative adverb.

Adjectival phrases extend beyond bare adjectives (ADJP → JJ) in a number of ways:

| | | |
|---|---|---|
| ADJP → RB JJ | *very hungry* | [8.65] |
| ADJP → RBR JJ | *more hungry* | [8.66] |
| ADJP → JJS JJ | *best possible* | [8.67] |
| ADJP → RB JJR | *even bigger* | [8.68] |
| ADJP → JJ CC JJ | *high and mighty* | [8.69] |
| ADJP → JJ JJ | *West German* | [8.70] |
| ADJP → RB VBN | *previously reported* | [8.71] |

The tags JJR and JJS refer to comparative and superlative adjectives respectively.

All of these phrase types can be coordinated:

| | | |
|---|---|---|
| PP →PP CC PP | *on time and under budget* | [8.72] |
| ADVP →ADVP CC ADVP | *now and two years ago* | [8.73] |
| ADJP →ADJP CC ADJP | *quaint and rather deceptive* | [8.74] |
| SBAR →SBAR CC SBAR | *whether they want control or whether they want exports* | [8.75] |

### 8.2.4 Grammatical ambiguity and weighted context-free grammars

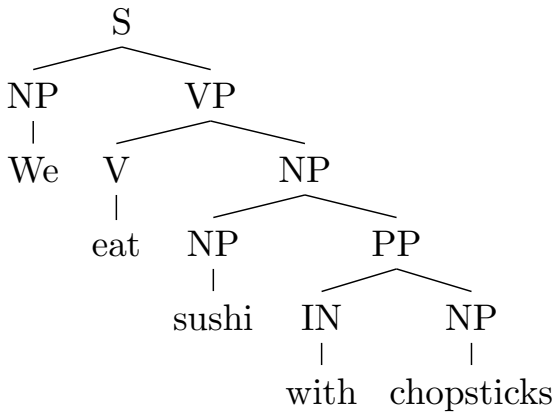
Context-free parsing is useful not only because it determines whether a sentence is grammatical, but mainly because it breaks the sentence up into constituents, which are useful in applications such as information extraction (chapter 16) and sentence compression (Jing, 2000; Clarke and Lapata, 2008). However, the **ambiguity** of wide-coverage natural language grammars poses a serious problem for such potential applications. As an example, Figure 8.13 shows two possible analyses for the simple sentence *she eats sushi with chopsticks*, depending on whether the *chopsticks* are modifying *eats* or *sushi*. Realistic grammars license thousands or even millions of parses for individual sentences.

Weighted context-free grammars solve this problem by attaching weights to each production. The score of a derivation $\tau = \{(X \rightarrow \alpha)\}$ is then the sum of the weights of the productions,

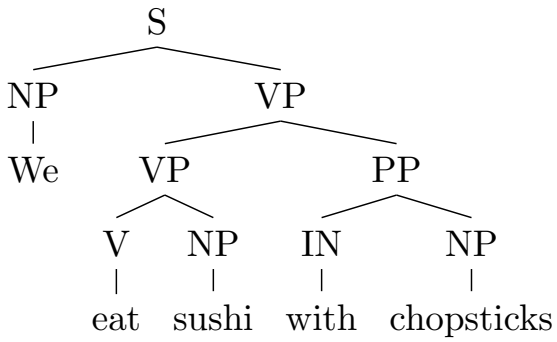
$$s(\tau) = \sum_{(X \rightarrow \alpha) \in \tau} \psi(X \rightarrow \alpha), \quad [8.76]$$

where $\psi(X \rightarrow \alpha)$ is the score of the production $X \rightarrow \alpha$. The parsing problem is then,

$$\hat{\tau} = \underset{\tau : \text{yield}(\tau) = \boldsymbol{w}}{\text{argmax}} \; s(\tau). \quad [8.77]$$




(S(NP(PRP *She*)(VP(VBZ *eats*)
(NP(NP(NN *sushi*))(PP (IN*with*)(NP(NNS *chopsticks*)))))))




(S(NP(PRP *She*)(VP(VBZ *eats*)
(NP(NN *sushi*))
(PP(IN*with*)(NP(NNS *chopsticks*))))))

Figure 8.13: Two derivations of the same sentence, shown as both parse trees and bracketings

Parsing algorithms for weighted CFGs are described in chapter 9; in general, the time complexity is cubic in the length of the input. Given a labeled dataset, the weights can be set equal to log conditional probabilities, $\psi(X \rightarrow \alpha) = \log \text{p}(\alpha \mid X)$. Setting the weights in this way defines a **probabilistic context-free grammar** (PCFG), in which the total score of a derivation is equal to the log of the joint probability, $\log \text{p}(\boldsymbol{w}, \tau)$. Alternatively, the weights can be learned using discriminative algorithms such as perceptron. For more details, see § 9.3.

## 8.3 *Mildly context-sensitive languages

Beyond context-free languages lie **context-sensitive languages**, in which the expansion of a non-terminal depends on its neighbors. In the general class of context-sensitive languages, computation becomes much more challenging: the membership problem for context-sensitive languages is PSPACE-complete. Since PSPACE contains the complexity class NP (problems that can be solved in polynomial time on a non-deterministic Turing machine), PSPACE-complete problems cannot be solved efficiently if P $\neq$ NP. Thus, designing an efficient parsing algorithm for the full class of context-sensitive languages is probably hopeless.[9]

However, Joshi (1985) identifies a set of properties that define **mildly context-sensitive** languages, which are a strict subset of context-sensitive languages. Like context-free languages, mildly context-sensitive languages are efficiently parseable. However, the mildly context-sensitive languages include non-context-free languages, such as the "copy language" $\{\boldsymbol{ww} \mid \boldsymbol{w} \in \Sigma^*\}$ and the language $a^m b^n c^m d^n$. Both are characterized by **cross-serial dependencies**, linking symbols at long distance across the string.[10] For example, in the language $a^n b^m c^n d^m$, each $a$ symbol is linked to exactly one $c$ symbol, regardless of the number of intervening $b$ symbols.

Shieber (1985) describes a fragment of Swiss-German, in which sentences such as *we let the children help Hans paint the house* are realized by listing all nouns before all verbs, i.e., *we the children Hans the house let help paint*. Furthermore, each noun's determiner is dictated by the noun's **case marking** (the role it plays with respect to the verb). Using an argument that is analogous to the earlier discussion of center-embedding (§ 8.2), Shieber argues that these case marking constraints are a cross-serial dependency, homomorphic to $a^m b^n c^m d^n$, and therefore not context-free.

[9] If PSPACE $\neq$ 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).

[10] A 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 $\{a^{2^n} \mid n \geq 0\}$.

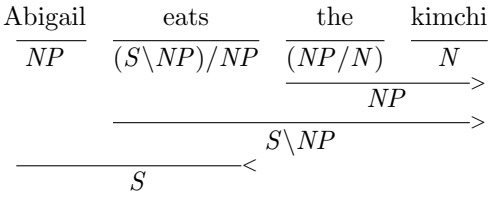



Figure 8.14: A syntactic analysis in CCG involving forward and backward function application

As with the move from regular to context-free languages, mildly context-sensitive languages can be motivated by expedience. While infinite sequences of cross-serial dependencies cannot be handled by context-free grammars, even finite sequences of cross-serial dependencies are more convenient to handle using a mildly context-sensitive formalism like **tree-adjoining grammar** (TAG) and **combinatory categorial grammar** (CCG). Furthermore, 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). Furthermore, 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 an overview of the two formalisms, but the reader is encouraged to consult Joshi and Schabes (1997) and Steedman and Baldridge (2011) for more detail on TAG and CCG respectively.

### 8.3.1 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, parenthese, 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\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\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 on the right, before the subject, which appears on the left. They therefore have the more com-

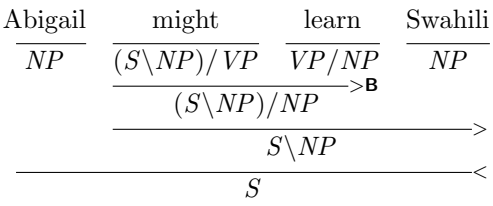



Figure 8.15: A syntactic analysis in CCG involving function composition (example based on Steedman and Baldridge, 2011)

plex type (S\NP)/NP. Similarly, the application of a determiner to the noun at its right results in a noun phrase, so determiners have the type NP/N. An example involving a transitive verb and a determiner is shown in Figure 8.14. A key point from this example is that it can be trivially transformed into phrase-structure tree, by treating each function application as a constituent phrase. Indeed, when CCG's only combinatory operators are forward and backward function application, it is equivalent to context-free grammar. However, the location of the "effort" has changed. Rather than designing good productions, the grammar designer must focus on the **lexicon** — choosing the right categories for each word. This makes it possible to parse a wide range of sentences using only a few generic combinatory operators.

Things become more interesting with the introduction of two additional operators: **composition** and **type-raising**. Function composition enables the combination of complex types: $X/Y \circ Y/Z \Rightarrow_{\mathbf{B}} X/Z$ (forward composition) and $Y\backslash Z \circ X\backslash Y \Rightarrow_{\mathbf{B}} X\backslash Z$ (backward composition).[11] Composition makes it possible to "look inside" complex types, and combine two adjacent units if the "input" for one is the "output" for the other. Figure 8.15 shows how function composition can be used to handle modal verbs. While this sentence can be parsed using only function application, the composition-based analysis is preferable because the unit *might learn* functions just like a transitive verb, as in the example *Abigail studies Swahili*. This in turn makes it possible to analyze conjunctions such as *Abigail studies and might learn Swahili*, attaching the direct object *Swahili* to the entire conjoined verb phrase *studies and might learn*.

Type raising converts an element of type $X$ to a more complex type: $X \Rightarrow_T T/(T\backslash X)$ (forward type-raising to type $T$), and $X \Rightarrow_T T\backslash(T/X)$ (backward type-raising to type $T$). Type-raising makes it possible to reverse the relationship between a function and its argument — by transforming the argument into a function over functions over arguments! An example may help demystify things. Figure 8.15 shows how to analyze an object relative clause, *a story that Abigail tells*. The problem is that *tells* is a transitive verb, expecting a direct object to its right. As a result, *Abigail tells* is not a constituent in our context-free grammar of English. The issue is resolved by raising *Abigail* from NP to the complex type

[11] The subscript $\mathbf{B}$ follows notation from Curry and Feys (1958).

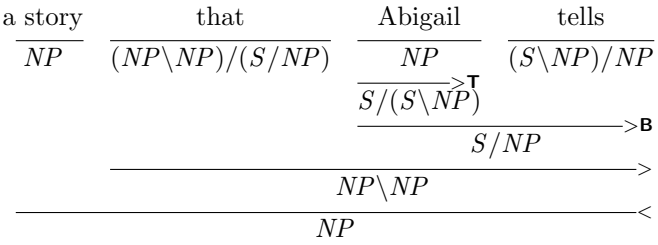



Figure 8.16: A syntactic analysis in CCG involving an object relative clause (based on slides from Alex Clark)

(S/NP)\NP). This function can then be combined with the transitive verb *tells* by forward composition, resulting in the type (S/NP), which is a sentence lacking a direct object to its right.[12] From here, we need only design the lexical entry for the complementizer *that* to expect a right neighbor of type (S/NP), and the remainder of the derivation can proceed by function application.

Composition and type-raising give CCG considerable power and flexibility, but at a price. The simple sentence *Abigail tells Max* can be parsed in two different ways: by function application (first forming the verb phrase *tells Max*), and by type-raising and composition (first forming the non-constituent *Abigail tells*). This **derivational ambiguity** does not affect the resulting linguistic analysis, so it is sometimes known as **spurious ambiguity**. Hockenmaier and Steedman (2007) present a translation algorithm for converting the Penn Treebank into CCG derivations, using composition and type-raising only when necessary.

### 8.3.2 Tree-adjoining grammar

[todo: todo]

## Exercises

1. Sketch out the state diagram for finite-state acceptors for the following languages on the alphabet $\{a, b\}$.

   a) Even-length strings. (Be sure to include 0 as an even number.)

   b) Strings that contain $aaa$ as a substring.

   c) Strings containing an even number of $a$ and an odd number of $b$ symbols.

[12] The missing direct object would be analyzed as a **trace** in CFG-like approaches to syntax, including the Penn Treebank.

   d) Strings in which the substring *bbb* must be terminal if it appears — the string need not contain *bbb*, but if it does, nothing can come after it.

2. Levenshtein edit distance is the number of insertions, substitutions, or deletions required to convert one string to another.

   a) Define a finite-state acceptor that accepts all strings with edit distance 1 from the target string, *target*.

   b) Now think about how to generalize your design to accept all strings with edit distance from the target string equal to $d$. If the target string has length $\ell$, what is the minimal number of states required?

3. Construct an FSA in the style of Figure 8.3, which handles the following examples:

   - *nation*/N, *national*/ADJ, *nationalize*/V, *nationalizer*/N
   - *America*/N, *American*/ADJ, *Americanize*/V, *Americanizer*/N

   Be sure that your FSA does not accept any further derivations, such as **nationalizeral* and **Americanizern*.

4. Show how to construct a trigram language model in a weighted finite-state acceptor. Make sure that you handle the edge cases at the beginning and end of the sequence accurately.

5. Extend the FST in Figure 8.6 to handle the other two parts of rule 1a of the Porter stemmer: *-sses* $\rightarrow$ *ss*, and *-ies* $\rightarrow$ *-i*.

6. § 8.1.4 describes $T_O$, a transducer that captures English orthography by transducing *cook* + *ed* $\rightarrow$ *cooked* and *bake* + *ed* $\rightarrow$ *baked*. Design an unweighted finite-state transducer that captures this property of English orthography.

   Next, augment the transducer to appropriately model the suffix *-s* when applied to words ending in *s*, e.g. *kiss+s* $\rightarrow$ *kisses*.

7. Add parenthesization to the grammar in Figure 8.11 so that it is no longer ambiguous.

8. Construct three examples — a noun phrase, a verb phrase, and a sentence — which can be derived from the Penn Treebank grammar fragment in § 8.2.3, yet are not grammatical. Avoid reusing examples from the text. Optionally, propose corrections to the grammar to avoid generating these cases.

9. Produce parses for the following sentences, using the Penn Treebank grammar fragment from § 8.2.3.

(8.21) This aggression will not stand.

(8.22) I can get you a toe.

(8.23) Sometimes you eat the bar and sometimes the bar eats you.

Then produce parses for three short sentences from a news article from this week.

10. * One advantage of CCG is its ability to handle coordination of both constituents and non-constituents:

(8.24) *Abigail and Max speak Swahili*

(8.25) *Abigail speaks and Max understands Swahili*

Define the lexical entry for *and* as

$$\textit{and} := (X/X)\backslash X, \quad [8.78]$$

where $X$ can refer to any type. Using this lexical entry, show how to parse the two examples above. In the second example, *Swahili* should be combined with the coordination *Abigail speaks and Max understands*, and not just with the verb *understands*.

# Chapter 9

# Context-free Parsing

Parsing is the task of determining whether a string can be derived from a given context-free grammar, and if so, how. The parse of a sentence is a hierarchical structure of nested constituents. This structure helps to answer the basic questions of who-did-what-to-whom, and is useful for various downstream tasks, such as semantic analysis and information extraction.

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:

$$\begin{aligned} \text{X} &\rightarrow \text{X X} \\ \text{X} &\rightarrow \textit{aardvark} \mid \textit{abacus} \mid \ldots \mid \textit{zyther} \end{aligned}$$

The second line indicates unary productions to every nonterminal in $\Sigma$. In this grammar, the number of possible derivations for a string $\boldsymbol{w}$ is equal to the number of binary bracketings, e.g.,

$$((((w_1 w_2) w_3) w_4) w_5), \quad (((w_1 (w_2 w_3)) w_4) w_5), \quad ((w_1 (w_2 (w_3 w_4))) w_5), \quad \ldots.$$

The number of binary bracketings is a **Catalan number**, which grows super-exponentially in the length of the sentence, $C_n = \frac{(2n)!}{(n+1)!n!}$. As with sequence labeling, it is only possible to search the space of possible derivations by resorting to independence assumptions, which allow us to search efficiently by reusing shared substructures with dynamic programming. This chapter will focus on a bottom-up dynamic programming algorithm called **CKY**. CKY enables exhaustive search of the space of possible parses, but imposes strict limitations on the form of scoring function. These limitations can be relaxed by abandoning exhaustive search. A few non-exact search methods will be discussed at the end of this chapter, and one of them — **transition-based parsing** — will be the focus of chapter 10.

| | |
|---|---|
| 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 9.1: A toy example context-free grammar

## 9.1 Deterministic bottom-up parsing

The **CKY algorithm**[1] is a bottom-up approach to parsing in a context-free grammar. It efficiently tests whether a string is in a language, without considering all possible parses. The algorithm first forms small constituents, and then tries to merge them into larger constituents.

To understand the algorithm, consider the input, *we eat sushi with chopsticks*. According to the toy grammar in Table 9.1, each terminal symbol can be generated by exactly one unary production, resulting in the sequence NP V NP IN NP. Next, we can try to apply binary productions to merge adjacent symbols into larger constituents: we could merge V NP into VP, or we could merge IN NP into PP. The goal of bottom-up parsing is to find some series of mergers that ultimately results in the start symbol S covering the entire input.

The CKY algorithm systematizes this approach, incrementally constructing a table (or **chart**) $t$ in which each cell $t[i, j]$ contains the set of nonterminals that can derive the span $\boldsymbol{w}_{i+1:j}$. The algorithm fills in the upper right triangle of the table; it begins with the diagonal, which corresponds to substrings of length 1, and the computes derivations for progressively larger substrings, until reaching the upper right corner $t[0, M]$, which corresponds to the entire input. If the start symbol S is in $t[0, M]$, then the string $\boldsymbol{w}$ is in the language defined by the grammar. This process is detailed in Algorithm 11, and the resulting data structure is shown in Figure 9.1. Informally, here's how it works:

- Begin by filling in the diagonal: the entries $t[m-1, m]$ for all $m \in \{1 \dots M\}$. These are filled with terminal productions that yield the individual tokens; for the word $w_2 = \textit{sushi}$, we fill in $t[1, 2] = \{\text{NP}\}$, and so on.
- Then fill in the next diagonal, in which each cell corresponds to a subsequence of length two: $t[0, 2], t[1, 3], \dots, t[M-2, M]$. These cells are filled in by looking for binary productions capable of producing at least one entry in each of the cells corre-

[1] The name is for Cocke-Kasami-Younger, the inventors of the algorithm. It is sometimes called **chart parsing**, because of its chart-like data structure.

**Algorithm 11** The CKY algorithm for parsing a sequence $\boldsymbol{w} \in \Sigma^*$ 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 (§ 8.2.1). The function 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 $\varnothing$.

```
1: procedure CKY(w, G = (N, Σ, R, S))
2:     for m ∈ {1 ... M} do
3:         t[m − 1, m] ← {X : (X → w_m) ∈ R}
4:     for ℓ ∈ {2 ... M} do                          ▷ Iterate over constituent lengths
5:         for m ∈ {0 ... M − ℓ} do                  ▷ Iterate over left endpoints
6:             for k ∈ {m + 1 ... m + ℓ − 1} do      ▷ Iterate over split points
7:                 for (X → Y Z) ∈ R do              ▷ Iterate over rules
8:                     if Y ∈ t[m, k] ∧ Z ∈ t[k, m + ℓ] then
9:                         t[m, m + ℓ] ← t[m, m + ℓ] ∪ X          ▷ Add non-terminal to table
10:                        b[m, m + ℓ, X] ← b[m, m + ℓ, X] ∪ (Y, Z, k)   ▷ Add back-pointers
11:    if S ∈ t[0, M] then
12:        return TRACEBACK(S, 0, M, b)
13:    else
14:        return ∅
15: procedure TRACEBACK(X, i, j, b)
16:    if j = i + 1 then
17:        return X
18:    else
19:        (Y, Z, k) ← PICKFROM(b[i, j, X])
20:        return X → (TRACEBACK(Y, i, k, t), TRACEBACK(Z, k, j, t))
```

sponding to left and right children. For example, the cell $t[1, 3]$ includes VP because the grammar includes the production VP → V NP, and 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 to make: where to split the left and right children. The cell $t[i, j]$ corresponds to the subsequence $\boldsymbol{w}_{i+1:j}$, and we must choose some *split point* $i < k < j$, so that $\boldsymbol{w}_{i+1:k}$ is the left child and $\boldsymbol{w}_{k+1:j}$ is the right child. We do this by considering all possible $k$, and 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 $t[k, j]$ have already been filled in, since these cells correspond to shorter sub-strings of the input.
- The process continues until we reach $t[0, M]$.

Figure 9.1 shows the chart that arises from parsing the sentence *we eat sushi with chopsticks* using the grammar defined above.

### 9.1.1 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 $X \rightarrow Y Z$ and the split point $k$, then we keep back-pointers $b[i, j, X] = (Y, Z, k)$. Once the table is constructed, we start with the back-pointers in $b[0, M, \text{S}]$, and recursively follow them until they ground out at terminal productions.

For ambiguous sentences, there will be multiple paths to reach $\text{S} \in t[0, M]$. For example, in Figure 9.1, the goal state $\text{S} \in t[0, M]$ is reached through the state $\text{VP} \in t[1, 5]$, 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 could have been generated by the grammar in more than one way. In Algorithm 11, one of these derivations is selected arbitrarily. § 9.3 describes how weighted context-free grammars can select the optimal parse, according to a scoring function.

### 9.1.2 Non-binary productions

The CKY algorithm assumes that all productions with non-terminals on the right-hand side (RHS) are binary. But in real grammars, such as the one considered in chapter 8, there will be productions with more than two elements on the right-hand side, and other productions with only a single element.

- For productions with more than two elements on the right-hand side, we **binarize**, creating additional non-terminals (see § 8.2.1). For example, if we have the production VP $\rightarrow$ V NP NP (for ditransitive verbs), we might convert to VP $\rightarrow$ $\text{VP}_{ditrans}$/NP NP, and then add the production $\text{VP}_{ditrans}/\text{NP} \rightarrow$ V NP.
- What about unary productions like VP $\rightarrow$ V? In practice, this is handled by making a second pass on each diagonal, in which each cell $t[i, j]$ is augmented with all possible unary productions capable of generating each item already in the cell — formally, $t[i, j]$ is extended to its **unary closure**. Suppose the example grammar in Table 9.1 were extended to include the production VP $\rightarrow$ V, enabling sentences with intransitive verb phrases, like *we eat*. Then the cell $t[1, 2]$ — corresponding to the word *eat* — would first include the set $\{\text{V}\}$, and would be augmented to the set $\{\text{V}, \text{VP}\}$ during this second pass.

### 9.1.3 Complexity

For an input of length $M$ and a grammar with $R$ productions and $N$ non-terminals, the space complexity of the CKY algorithm is $\mathcal{O}(M^2 N)$: the number of cells in the chart is $\mathcal{O}(M^2)$, and each cell must hold $\mathcal{O}(N)$ elements. The time complexity is $\mathcal{O}(M^3 R)$. Each

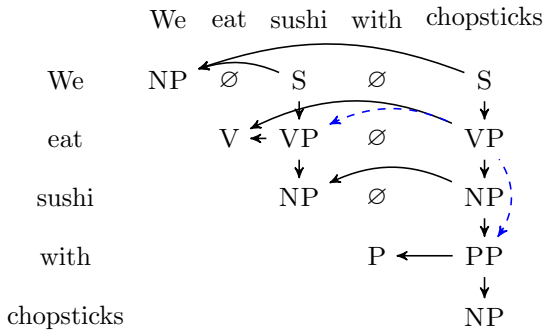



Figure 9.1: An example completed CKY chart. The solid and dashed lines show the back pointers resulting from the two different derivations of VP in position $t[1, 5]$.

cell is computed by searching over $\mathcal{O}(M)$ split points, with $R$ possible productions for each split point. Both the time and space complexity are considerably worse than the finite-state algorithms of Viterbi and forward-backward, which are linear in the length of the input.

## 9.2 Ambiguity

Syntactic ambiguity is endemic to natural language. Here are a few broad categories:

- **Attachment ambiguity**: e.g., *we eat sushi with chopsticks*, *I shot an elephant in my pajamas*. In each of these examples, the prepositions (*with*, *in*) can attach to either the verb or the direct object.
- **Modifier scope**: e.g., *southern food store*, *plastic cup holder*. In these examples, the first word could be modifying the subsequent adjective, or the final noun.
- **Particle versus preposition**: e.g., *the puppy tore up the staircase*. Phrasal verbs like *tore up* often include particles which could also act as prepositions.
- **Complement structure**: e.g., *the students complained to the professor that they didn't understand*. This is another form of attachment ambiguity, where the complement *that they didn't understand* could attach to the main verb (*complained*), or to the indirect object (*the professor*).
- **Coordination scope**: e.g., *"I see," said the blind man, as he picked up the hammer and saw*. In this example, the lexical ambiguity for *saw* enables it to be coordinated either with the noun *hammer* or the verb *picked up*.

These forms of ambiguity can combine, so that seemingly simple headlines like *Fed raises interest rates* have dozens of possible analyses, even in a minimal grammar. Broad

coverage grammars permit millions of parses of typical sentences. While careful grammar design can chip away at this ambiguity, a better strategy is combine broad-coverage parsers with data-driven strategies for identifying the correct analysis.

### 9.2.1 Parser evaluation

Before continuing to parsing algorithms that are able to handle ambiguity, we stop to consider how to measure parsing performance. Suppose we have a set of *reference parses* — the ground truth — and a set of *system parses* that we would like to score. A simple solution would be per-sentence accuracy: the parser is scored by the proportion of sentences on which the system and reference parses exactly match.[2] But as any good student knows, it is better to get *partial credit*, which we can assign to analyses that correctly match parts of the reference parse. The PARSEval metrics (Grishman et al., 1992) score each system parse via:

**Precision:** the fraction of brackets in the system parse that match a bracket in the reference parse.

**Recall:** the fraction of brackets in the reference parse that match a bracket in the system parse.

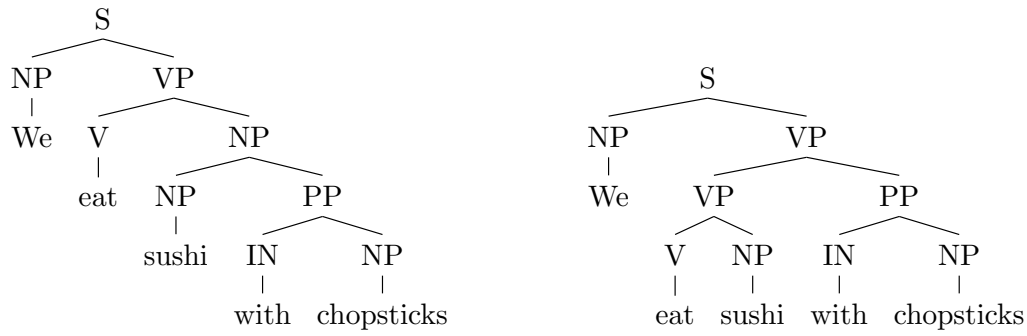
In **labeled precision** and **recall**, the system must also match the non-terminals for each bracket; in **unlabeled precision** and **recall**, it is only required to match the bracketing structure. As in chapter 3, the precision and recall can be combined into an $F$-measure, $F = \frac{2 \times P \times R}{P+R}$.

In Figure 9.2, suppose that the left tree is the system parse and the right tree is the reference parse. We have the following spans:

- S $\rightarrow \boldsymbol{w}_{1:5}$ is *true positive*, because it appears in both trees.
- VP $\rightarrow \boldsymbol{w}_{2:5}$ is *true positive* as well.
- NP $\rightarrow \boldsymbol{w}_{3:5}$ is *false positive*, because it appears only in the system output.
- PP $\rightarrow \boldsymbol{w}_{4:5}$ is *true positive*, because it appears in both trees.
- VP $\rightarrow \boldsymbol{w}_{2:3}$ is *false negative*, because it appears only in the reference.

The labeled and unlabeled precision of this parse is $\frac{3}{4} = 0.75$, and the recall is $\frac{3}{4} = 0.75$, for an F-measure of $0.75$. For an example in which precision and recall are not equal, suppose the reference parse instead included the production VP $\rightarrow$ V NP PP. In this parse, the reference does not contain the constituent $\boldsymbol{w}_{2:3}$, so the recall would be $1$.

[2] Most parsing papers do not report results on this metric, but Finkel et al. (2008) find that a near-state-of-the-art parser finds the exact correct parse on 35% of sentences of length $\leq 40$, and on 62% of parses of length $\leq 15$ in the Penn Treebank.[todo: update]




(a) system output (b) reference

Figure 9.2: Two possible analyses from the grammar in Table 9.1

### 9.2.2 Local solutions

Some ambiguity can be resolved locally. Consider the following examples,

(9.1) We met the President on Monday.

(9.2) We met the President of Mexico.

Each case ends with a preposition, which can be attached to the verb *met* or the noun phrase *the president*. To resolve this ambiguity, we can use a labeled corpus to compute the likelihood of the observing the preposition alongside each candidate attachment point,

$$\text{p}(\textit{on} \mid \textit{met}), \quad \text{p}(\textit{on} \mid \textit{President}) \tag{9.1}$$

$$\text{p}(\textit{of} \mid \textit{met}), \quad \text{p}(\textit{of} \mid \textit{President}). \tag{9.2}$$

A comparison of these probabilities would successfully resolve this case (Hindle and Rooth, 1993). Other cases, such as the example *. . . eat sushi with chopsticks*, require considering the object of the preposition as well. With sufficient labeled data, the problem of prepositional phrase attachment can be treated as a text classification task (Ratnaparkhi et al., 1994).

However, there are inherent limitations to local solutions. While toy examples may have just a few ambiguities to resolve, realistic sentences have thousands or millions of possible parses. Furthermore, attachment decisions are interdependent, as shown in the following garden path example:

(9.3) *Cats scratch people with claws with knives.*

We may want to attach *with claws* to *scratch*, as would be correct in the shorter sentence in *cats scratch people with claws*. But this leaves nowhere to attach *with knives*. The correct

interpretation can be identified only be considering the attachment decisions jointly. The huge number of potential parses may seem to make exhaustive search impossible. But as with sequence labeling, we can use independence assumptions to search this space efficiently.

## 9.3 Weighted Context-Free Grammars

In a **weighted context-free grammar** (WCFG), each production $X \rightarrow \alpha$ is associated with a score $\psi(X \rightarrow \alpha)$. The score of a derivation is $\tau$ the combination of the scores of all the productions:

$$\Psi(\tau) = \bigotimes_{(X \rightarrow \alpha, (i,j,k)) \in \tau} \psi(X \rightarrow \alpha), \quad [9.3]$$

where $\tau$ is regarded as a set of **anchored productions** $\tau = \{X \rightarrow \alpha, (i, j, k)\}$, with $X$ corresponding to the left-hand side non-terminal, $\alpha$ corresponding to the right-hand side; for grammars in Chomsky normal formal, $\alpha$ is either a pair of non-terminals or a terminal symbol. The tuple $(i, j, k)$ anchors the production in the input, with $X$ deriving the span $\boldsymbol{w}_{i+1:j}$. For binary productions, $\boldsymbol{w}_{i+1:k}$ indicates the span of the left child, and $\boldsymbol{w}_{k+1:j}$ indicates the span of the right child; for unary productions, $k$ is ignored. In a weighted context-free grammar, the anchors are not considered in the production scores $\psi(X \rightarrow \alpha)$, but they are required to link together the productions into an analysis of the input.

The production scores are combined with the semiring operator $\bigotimes$ (see § 6.4.4), which unifies several types of weighted context free grammars as special cases.

- When each $\psi(X \rightarrow \alpha)$ is a *conditional probability* $\text{p}(\alpha \mid X)$, then $\bigotimes$ corresponds to multiplication, so that $\Psi(\tau)$ is the probability of the full derivation. This special case is explored in § 9.3.2.
- When $\psi(X \rightarrow \alpha)$ is a *conditional log-probability*, $\bigotimes$ corresponds to addition, so that $\Psi(\tau)$ is the log probability of the full derivation. We also set $\bigotimes$ to addition when each $\psi(X \rightarrow \alpha)$ is the output of a discriminatively-trained scoring function, such as $\psi(X \rightarrow \alpha) = \boldsymbol{\theta} \cdot \boldsymbol{f}(X, \alpha, \boldsymbol{w})$.
- When $\psi(X \rightarrow \alpha)$ is a *Boolean truth value* $\{\top, \bot\}$, then $\bigotimes$ is logical conjunction. This is equivalent to the unweighted context-free grammar considered in the previous section.

Regardless of how the scores are combined, the key point is the independence assumption: the score for a derivation is the combination of the independent scores for each production, and these scores do not depend on any other part of the derivation. For example, if two non-terminals are siblings, the scores of productions from these non-terminals

| | | $\psi(\cdot)$ | $\exp \psi(\cdot)$ |
|---|---|---|---|
| S | $\rightarrow$ NP VP | 0 | 1 |
| NP | $\rightarrow$ NP PP | $-1$ | $\frac{1}{2}$ |
| | $\rightarrow$ *we* | $-2$ | $\frac{1}{4}$ |
| | $\rightarrow$ *sushi* | $-3$ | $\frac{1}{8}$ |
| | $\rightarrow$ *chopsticks* | $-3$ | $\frac{1}{8}$ |
| PP | $\rightarrow$ IN NP | 0 | 1 |
| IN | $\rightarrow$ *with* | 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 | $\rightarrow$ *eat* | 0 | 1 |

Table 9.2: 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.

are computed independently. This independence assumption is analogous to the first-order Markov assumption in sequence labeling, where the score for transitions between tags depends only on the previous tag and current tag, and not on the history. As with sequence labeling, this assumption makes it possible to find the optimal parse efficiently; its linguistic limitations are explored in § 9.5.

**Example** Consider the weighted grammar shown in Table 9.2, and the analysis in Figure 9.2b. Since the weights are not probabilites, we $\otimes = +$, obtaining the following score:

$$\begin{aligned}\Psi(\tau) =& \psi(\text{S} \rightarrow \text{NP VP}) + \psi(\text{VP} \rightarrow \text{VP PP}) + \psi(\text{VP} \rightarrow \text{V NP}) + \psi(\text{PP} \rightarrow \text{IN NP}) \\ &+ \psi(\text{NP} \rightarrow \textit{we}) + \psi(\text{V} \rightarrow \textit{eat}) + \psi(\text{NP} \rightarrow \textit{sushi}) + \psi(\text{IN} \rightarrow \textit{with}) + \psi(\text{NP} \rightarrow \textit{chopsticks}) \qquad [9.4] \\ =& 0 - 2 - 1 + 0 - 2 + 0 - 3 + 0 - 3 = -11. \qquad [9.5]\end{aligned}$$

In the alternative parse in Figure 9.2a, the production VP $\rightarrow$ VP PP (with score $-2$) is replaced with the production NP $\rightarrow$ NP PP (with score $-1$); all other productions are the same. As a result, the score for this parse is $-10$.

This example hints at a big problem with WCFG parsing on non-terminals such as NP, VP, and PP: a WCFG will *always* prefer either VP or NP attachment, without regard to what is being attached! This problem is addressed in § 9.5.

### 9.3.1 Parsing with weighted context-free grammars

In a weighted context-free grammar, the task of parsing is to identify the best-scoring derivation,

$$\hat{\tau} = \underset{\tau:\text{yield}_G(\tau)=\boldsymbol{w}}{\text{argmax}} \bigotimes_{(X\rightarrow\alpha,(i,j,k))\in\tau} \psi(X \rightarrow \alpha),$$

The set $\{\tau : \boldsymbol{w} = \text{yield}_G(\tau)\}$ is the set of all possible derivations of $\boldsymbol{w}$ in a weighted context-free grammar $G$.

This optimization problem can be solved by modifying the CKY algorithm. In the deterministic CKY algorithm, each cell $t[i, j]$ stored a set of non-terminals capable of deriving the span $\boldsymbol{w}_{i+1:j}$. We now augment the table so that the cell $t[i, j, X]$ is the *score of the best-scoring derivation* of $w_{i+1:j}$ from non-terminal $X$. This score is computed recursively: for anchored binary production $(X \rightarrow Y\ Z, (i, j, k))$, we compute:

- the score of the production, $\psi(X \rightarrow Y\ Z)$;
- the score of the left child, $t[i, k, Y]$;
- the score of the right child, $t[k, j, Z]$.

These scores are combined by semiring multiplication $\otimes$. As in the unscored CKY algorithm, the table is constructed by considering spans of increasing length, so the scores for spans $t[i, k, Y]$ and $t[k, j, Z]$ are guaranteed to be available at the time we compute the score $t[i, j, X]$. The value $t[0, M, \text{S}]$ is the score of the best derivation of $\boldsymbol{w}$ from the grammar. Algorithm 12 formalizes this procedure.

As in unweighted CKY, the parse is recovered from the table of back pointers $b$, where each $b[i, j, X]$ stores the argmax split point $k$ and production $X \rightarrow Y\ Z$ in the derivation of $\boldsymbol{w}_{i+1:j}$ from $X$. The best parse can be obtained by tracing these pointers backwards from $b[0, M, \text{S}]$, 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 9.1. 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. At the next diagonal, we compute 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. In the example, each of these spans is unambiguous. Things get interesting when we reach

**Algorithm 12** CKY algorithm for parsing a string $\boldsymbol{w} \in \Sigma^*$ in a weighted context-free grammar $(N, \Sigma, R, S)$, 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 (§ 8.2.1). The function TRACEBACK is defined in Algorithm 11.

**procedure** WCKY($\boldsymbol{w}, G = (N, \Sigma, R, S)$)
  **for all** $i, j, X$ **do** ▷ Initialization
    $t[i, j, X] \leftarrow \overline{1}$
    $b[i, j, X] \leftarrow \varnothing$
  **for** $m \in \{1, \ldots, M\}$ **do**
    **for all** $X \in N$ **do**
      $t[m, m+1, X] \leftarrow \psi(X \rightarrow w_m)$
  **for** $\ell \in \{2 \ldots M\}$ **do**
    **for** $m \in \{0, \ldots, M - \ell\}$ **do**
      **for** $k \in \{m+1, \ldots, m+\ell-1\}$ **do**
        $t[m, m+\ell, X] \leftarrow \max_{k,Y,Z} \psi(X \rightarrow Y\ Z) \otimes t[m, k, Y] \otimes t[k, m+\ell, Z]$
        $b[m, m+\ell, X] \leftarrow \operatorname{argmax}_{k,Y,Z} \psi(X \rightarrow Y\ Z) \otimes t[m, k, Y] \otimes t[k, m+\ell, Z]$
  **return** TRACEBACK($S, 0, M, b$)

the cell $t[1, 5, \text{VP}]$, which contains the score for the derivation of the span $\boldsymbol{w}_{2:5}$ from the non-terminal VP. This score is computed as a max over two alternatives,

$$\begin{aligned} t[1, 5, \text{VP}] &= \max(\psi(\text{VP} \rightarrow \text{VP PP}) + t[1, 3, \text{VP}] + t[3, 5, \text{PP}], \\ &\qquad \psi(\text{VP} \rightarrow \text{V NP}) + t[1, 2, \text{V}] + t[2, 5, \text{NP}]) \qquad [9.6] \\ &= \max(-2 - 4 - 3, -1 + 0 - 7) = -8. \qquad [9.7] \end{aligned}$$

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.

### 9.3.2 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 $\psi(X \rightarrow \alpha) = \text{p}(\alpha \mid X)$, where the probability of the right-hand side $\alpha$ is conditioned on the non-terminal $X$. These probabilities must be normalized over all possible right-hand sides, so that $\sum_\alpha \text{p}(\alpha \mid X) = 1$, for all $X$. For a given parse $\tau$, the product of the probabilities of the productions is equal to $\text{p}(\tau)$, under the **generative model** $\tau \sim$ DRAWSUBTREE($S$), where the function DRAWSUBTREE is defined in Algorithm 13.

The conditional probability of a parse given a string is,

$$\text{p}(\tau \mid \boldsymbol{w}) = \frac{\text{p}(\tau)}{\sum_{\tau' : \text{yield}(\tau') = \boldsymbol{w}} \text{p}(\tau')}. \qquad [9.8]$$

**Algorithm 13** Generative model for derivations from probabilistic context-free grammars in Chomsky Normal Form (CNF).

```
procedure DRAWSUBTREE(X)
    sample (X → α) ∼ p(α | X)
    if α = (Y Z) then
        return DRAWSUBTREE(Y) ∪ DRAWSUBTREE(Z)
    else
        return (X → α)          ▷ In CNF, all unary productions yield terminal symbols
```

The numerator $\text{p}(\tau)$ can be computed by the CKY algorithm for weighted context-free grammars, where $\otimes = \times$. As a result, the CKY algorithm can identify $\hat{\tau} = \text{argmax}_\tau \, \text{p}(\tau)$, without computing the normalization term in the denominator. If a normalized probability $\text{p}(\tau \mid \boldsymbol{w})$ is required, the denominator of Equation 9.8 can be computed by the **inside recurrence**, described below.

**Example** The WCFG in Table 9.2 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 that yield the string *we eat sushi with chopsticks*, $\Psi(\tau_1) = \log \text{p}(\tau_1) = -10$ and $\Psi(\tau_2) = \log \text{p}(\tau_2) = -11$. Therefore, the conditional probability $\text{p}(\tau_1 \mid \boldsymbol{w})$ is equal to,

$$\text{p}(\tau_1 \mid \boldsymbol{w}) = \frac{\text{p}(\tau_1)}{\text{p}(\tau_1) + \text{p}(\tau_2)} = \frac{\exp \Psi(\tau_1)}{\exp \Psi(\tau_1) + \exp \Psi(\tau_2)} = \frac{2^{-10}}{2^{-10} + 2^{-11}} = \frac{2}{3}. \quad [9.9]$$

**The inside recurrence** The denominator of Equation 9.8 can be viewed as a language model, summing over all parses that yield the string $\boldsymbol{w}$,

$$\text{p}(\boldsymbol{w}) = \sum_{\tau' : \text{yield}\tau' = \boldsymbol{w}} \text{p}(\tau'). \quad [9.10]$$

Just as the CKY algorithm makes it possible to maximize over all such analyses, with a slight modification it can also compute their sum. The only change that is required is to replace the maximization over split points $k$ and productions $X \rightarrow Y \; Z$ with a sum. The two algorithms can be written in a single general form using the semiring addition operator $\oplus$: to find the best parse, we set $\oplus$ to maximization; to compute the sum of parses, we set $\oplus$ to addition. Thus, the cell $t[i, j, X]$ now contains the total probability of

deriving $\boldsymbol{w}_{i+1:j}$ from $X$,

$$t[i,j,X] = \bigoplus_{k,Y,Z} \psi(X \rightarrow Y\ Z) \otimes t[i,k,Y] \otimes t[k,j,Z] \quad [9.11]$$

$$= \sum_{k,Y,Z} \text{p}(Y\ Z \mid X) \times \text{p}(Y \rightarrow \boldsymbol{w}_{i+1:k}) \times \text{p}(Z \rightarrow \boldsymbol{w}_{k+1:j}) \quad [9.12]$$

$$= \text{p}(X \rightarrow \boldsymbol{w}_{i+1:j}). \quad [9.13]$$

The cell $t[0, M, S]$ is then the probability of deriving the entire input, $\text{p}(S \rightarrow \boldsymbol{w})$. The name "inside recurrence" implies a corresponding **outside recurrence**, which computes the probability of a non-terminal $X$ spanning $\boldsymbol{w}_{i+1:j}$, joint with the outside context $(\boldsymbol{w}_{1:i}, \boldsymbol{w}_{j+1:M})$. This recurrence is described in § 9.4.3.

## 9.4 Learning weighted context-free grammars

Like sequence labeling, context-free parsing is a form of structure prediction. As a result, WCFGs can be learned using the same set of algorithms: generative probabilistic models, structured perceptron, maximum conditional likelihood, and maximum margin learning. In all cases, learning requires a **treebank**, which is a dataset of sentences labeled with context-free parses. Parsing research was catalyzed by the **Penn Treebank** (Marcus et al., 1993), the first large-scale dataset of this type (see § 8.2.2). Phrase structure treebanks exist for roughly two dozen other languages, with coverage mainly restricted to European and East Asian languages, plus Arabic and Urdu.

### 9.4.1 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:

$$\psi(X \rightarrow \alpha) = \text{p}(X \rightarrow \alpha) \quad [9.14]$$

$$\hat{\text{p}}(X \rightarrow \alpha) = \frac{\text{count}(X \rightarrow \alpha)}{\text{count}(X)}. \quad [9.15]$$

For example, the probability of the production NP $\rightarrow$ DET NN is the corpus count of this production divided by the count of the non-terminal NP. This estimator applies to terminal productions as well: the probability of NN $\rightarrow$ *whale* is the count of how often *whale* appears in the corpus as generated from an NN tag, divided by the total count of the NN tag. Even with the largest treebanks — currently on the order of one million tokens — it is difficult to accurately compute probabilities of even moderately rare events, such as NN $\rightarrow$ *whale*. Therefore, smoothing is critical for making PCFGs effective.

### 9.4.2 Feature-based parsing

The scores for each production can be computed as an inner product of weights and features,

$$\psi(X \rightarrow \alpha) = \boldsymbol{\theta} \cdot \boldsymbol{f}(X, \alpha), \tag{9.16}$$

where the feature vector $\boldsymbol{f}(X, \alpha)$ is a function of the left-hand side $X$ and the right-hand side $\alpha$. The basic feature $\boldsymbol{f}(X, \alpha) = \{(X, \alpha)\}$ encodes only the identity of the production itself, which is a discriminatively-trained model with the same expressiveness as a PCFG. Other features can be obtained by grouping elements on either the left-hand or right-hand side. It can be particularly beneficial to compute additional features by clustering terminal symbols, with features corresponding to groups of words with similar syntactic properties. The clustering can be obtained from unlabeled datasets that are much larger than any treebank, improving coverage. Such methods are described in chapter 13.

It is also possible to build features that include lexical information about the span and its boundaries. Such features are defined over **anchored productions**,

$$\psi(X \rightarrow \alpha, (i, j, k)) = \boldsymbol{\theta} \cdot \boldsymbol{f}(X, \alpha, (i, j, k), \boldsymbol{w})$$

where the feature vector is now a function of the details of the production ($X$ and $\alpha$), as well as the text $\boldsymbol{w}$ and the indices of the spans to derive: the parent $\boldsymbol{w}_{i+1:j}$, the left child $\boldsymbol{w}_{i+1:k}$, and the right child $\boldsymbol{w}_{k+1:j}$. Features on anchored productions can include the words that border the 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 child span $w_{i+1:k}$, and so on (Durrett and Klein, 2015). Strictly speaking, the WCFG formalism scores only the productions $X \rightarrow \alpha$, and not anchored productions. However, scores on anchored productions can be incorporated into CKY parsing without any modification to the algorithm, because it is still possible to compute each element of the table $t[i, j, X]$ recursively from its immediate children.

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),

$$\boldsymbol{f}(\tau, \boldsymbol{w}^{(i)}) = \sum_{(X \rightarrow \alpha, (i,j,k)) \in \tau} \boldsymbol{f}(X, \alpha, (i, j, k), \boldsymbol{w}^{(i)}) \tag{9.17}$$

$$\hat{\tau} = \underset{\tau : \text{yield}(\tau) = \boldsymbol{w}^{(i)}}{\text{argmax}} \; \boldsymbol{\theta} \cdot \boldsymbol{f}(\tau, \boldsymbol{w}^{(i)}) \tag{9.18}$$

$$\boldsymbol{\theta} \leftarrow \boldsymbol{f}(\tau^{(i)}, \boldsymbol{w}^{(i)}) - \boldsymbol{f}(\hat{\tau}, \boldsymbol{w}^{(i)}). \tag{9.19}$$

A margin-based objective can be optimized by selecting $\hat{\tau}$ through cost-augmented decoding (§ 1.3.2), enforcing a margin of $\Delta(\hat{\tau}, \tau)$ between the hypothesis and the reference parse, where $\Delta$ is a non-negative cost function, such as the Hamming loss (Stern et al., 2017). It is also possible to train feature-based parsing models by conditional log-likelihood, as described next.

### 9.4.3 *Conditional random field parsing

The score of a derivation $\Psi(\tau)$ can be converted into a probability by normalizing over all possible derivations,

$$\mathrm{p}(\tau \mid \boldsymbol{w}) = \frac{\exp \Psi(\tau)}{\sum_{\tau': \text{yield}(\tau')=\boldsymbol{w}} \exp \Psi(\tau')}. \quad [9.20]$$

This suggests another possible objective for learning a WCFG, maximizing the conditional log-likelihood of a labeled corpus.

Just as in logistic regression and the sequence CRF, the gradient of the conditional log-likelihood is the difference between the observed and expected counts of each feature. The expectation $E_{\tau|\boldsymbol{w}}[\boldsymbol{f}(\tau, \boldsymbol{w}^{(i)}); \boldsymbol{\theta}]$ requires summing over all possible parses, and computing the marginal probabilities of anchored productions, $\mathrm{p}(X \rightarrow \alpha, (i, j, k) \mid \boldsymbol{w})$. In CRF sequence labeling, similar marginal probabilities are computed by the two-pass **forward-backward algorithm** (§ 6.5.3). The analogue for context-free grammars is the **inside-outside algorithm**, in which marginal probabilities are computed from terms generated by an upward and downward pass over the parsing chart:

- The upward pass is performed by the **inside recurrence**, which is described in § 9.3.2. It computes the score,

$$\alpha(i, j, X) \triangleq \bigoplus_{(X \rightarrow Y\ Z)} \bigoplus_{k=i}^{j} \psi(X \rightarrow Y\ Z, (i, j, k)) \otimes \alpha(i, k, Y) \otimes \alpha(k, j, Z). \quad [9.21]$$

  The initial condition of this recurrence is $\alpha(i-1, i, X) = \psi(X \rightarrow w_i)$. To compute log-probabilities, the recurrence is applied in a semiring where $a \otimes b = a + b$ and $a \oplus b = \log(\exp x + \exp y)$. Thus, if $\alpha(i, k, Y)$, $\alpha(k, j, Z)$, and $\Psi(X \rightarrow Y\ Z)$ are log-probabilities, then $\alpha(i, j, X)$ will be a log-probability as well. The denominator $\sum_{\tau: \text{yield}(\tau)=\boldsymbol{w}} \exp \Psi(\tau)$ is equal to $\exp \alpha(0, M, \mathrm{S})$.

- The downward pass is performed by the **outside recurrence**, which recursively populates the same chart structure, starting at the root of the tree. It computes the probability,

$$\beta(i, j, X) \triangleq \left( \bigoplus_{(Y \rightarrow Z\ X)} \bigoplus_{k=0}^{i} \psi(Y \rightarrow Z\ X, (k, i, j)) \otimes \alpha(k, i, Z) \otimes \beta(k, j, Y) \right) \quad [9.22]$$

$$\oplus \left( \bigoplus_{(Y \rightarrow X\ Z)} \bigoplus_{k=j+1}^{M} \Psi(Y \rightarrow X\ Z, (i, k, j)) \otimes \alpha(j+1, k, Z) \otimes \beta(i, k, Y) \right). \quad [9.23]$$

[todo: check for off-by-one errors] The first line of Equation 9.23 is the score under the condition that $X$ is a right child of its parent $Y$, which spans $\boldsymbol{w}_{k:j}$, with $k < i$; the second line is the score under the condition that $X$ is a left child of its parent, which spans $\boldsymbol{w}_{i:k}$, with $k > j$. In each case, we sum over all possible productions with $X$ on the right-hand side. The parent $Y$ is bounded on one side by either $i$ or $j$, depending on whether $X$ is a left or right child of $Y$; we must sum over all possible values for the other boundary.

The initial condition for the outside recurrence is $\beta(1, M, \text{S}) = \overline{1}$, where $\overline{1}$ is the multiplicative identity, $x \otimes \overline{1} = x$, and $\beta(1, M, X \neq \text{S}) = \overline{0}$, where $\overline{0}$ is the multiplicative annihilator, $x \otimes \overline{0} = \overline{0}$. In the log-probability semiring, $\overline{1} = 0$, and $\overline{0} = -\infty$.

The marginal probability of a non-terminal $X$ over span $\boldsymbol{w}_{i:j}$ is written $\text{p}(X \rightsquigarrow \boldsymbol{w}_{i:j} \mid \boldsymbol{w})$, and depends on the sum of the scores of all parses $\tau$ that derive the span $\boldsymbol{w}_{i:j}$ from the non-terminal $X$, which is computed as $\alpha(i, j, X) \otimes \beta(i, j, X)$. This can be converted to a probability by exponentiating and normalizing,

$$\text{p}(X \rightsquigarrow \boldsymbol{w}_{i:j} \mid \boldsymbol{w}_{1:i-1}, \boldsymbol{w}_{j+1:M}) = \frac{\exp(\alpha(i, j, X) \otimes \beta(i, j, X))}{\exp(\alpha(1, M, S))}. \quad [9.24]$$

Probabilities of individual productions can be computed similarly (see exercises). The same marginal probabilities can be used in unsupervised **grammar induction**, in which a PCFG is learned from a dataset of unlabeled text (Lari and Young, 1990; Pereira and Schabes, 1992).

### 9.4.4 Neural context-free grammars

Recent work has applied neural representations to parsing, representing each span with a dense numerical vectors (Socher et al., 2013; Durrett and Klein, 2015; Cross and Huang, 2016). For example, the anchor $(i, j, k)$ and sentence $\boldsymbol{w}$ can be associated with a fixed-length column vector,

$$\boldsymbol{v}_{(i,j,k)} = [\boldsymbol{u}_{w_{i-1}}; \boldsymbol{u}_{w_i}; \boldsymbol{u}_{w_{j-1}}; \boldsymbol{u}_{w_j}; \boldsymbol{u}_{w_{k-1}}; \boldsymbol{u}_{w_k}], \quad [9.25]$$

where $\boldsymbol{u}_{w_i}$ is a word embedding associated with the word $w_i$. The vector $\boldsymbol{v}_{i,j,k}$ 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),

$$\tilde{\boldsymbol{v}}_{(i,j,k)} = \text{FeedForward}(\boldsymbol{v}_{(i,j,k)}) \quad [9.26]$$

$$\psi(X \rightarrow \alpha, (i, j, k)) = \tilde{\boldsymbol{v}}_{(i,j,k)}^\top \boldsymbol{\Theta} \boldsymbol{f}(X \rightarrow \alpha), \quad [9.27]$$

where $\boldsymbol{f}(X \rightarrow \alpha)$ is a vector of discrete features of the production, and $\boldsymbol{\Theta}$ is a parameter matrix. The matrix $\boldsymbol{\Theta}$ and the parameters of the feedforward network can be learned by

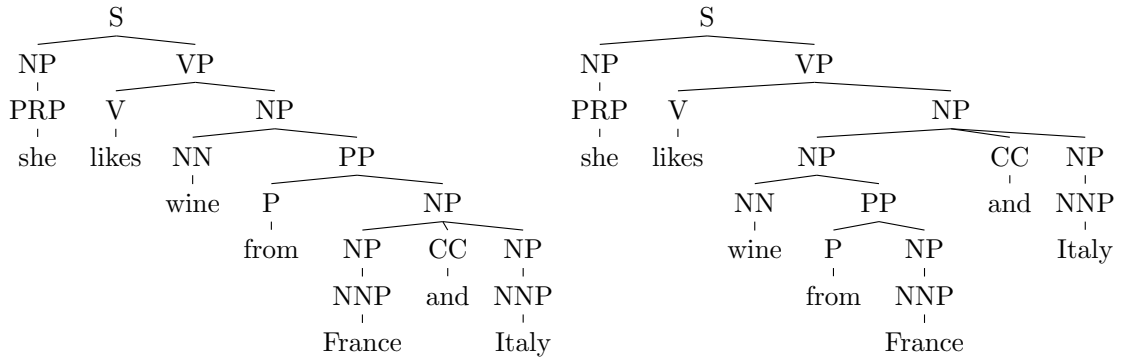



Figure 9.3: The left parse is preferable because of the conjunction of phrases headed by *France* and *Italy*, but these parses cannot be distinguished by a WCFG.

backpropagating from an objective such as the margin loss or the negative conditional log-likelihood. [todo: henderson and titov?]

## 9.5 Grammar refinement

The strength of the independence assumptions underlying CFG parsing depends on the granularity of the non-terminals. For the Penn Treebank non-terminals, there are several reasons to believe that these assumptions are too strong to enable accurate parsing (Johnson, 1998):

- The context-free assumption is too strict: for example, the probability of the production NP $\rightarrow$ NP PP is much higher (in the PTB) if the parent of the noun phrase is a verb phrase (indicating that the NP is a direct object) than if the parent is a sentence (indicating that the NP is the subject of the sentence).
- The Penn Treebank non-terminals are too coarse: there are many kinds of noun phrases and verb phrases, and accurate parsing sometimes requires knowing the difference. As we have already seen, when faced with prepositional phrase attachment ambiguity, a weighted CFG will either always choose NP attachment (if $\psi(\text{NP} \rightarrow \text{NP PP}) > \psi(\text{VP} \rightarrow \text{VP PP})$), or it will always choose VP attachment. To get more nuanced behavior, more fine-grained non-terminals are needed.
- More generally, accurate parsing requires some amount of **semantics** — understanding the meaning of the text to be parsed. Consider the example *cats scratch people with claws*: knowledge of about *cats*, *claws*, and scratching is necessary to correctly resolve the attachment ambiguity.

An extreme example is shown in Figure 9.3. The analysis on the left is preferred because of the conjunction of similar entities *France* and *Italy*. But given the non-terminals shown in the analyses, there is no way to differentiate these two parses, since they include exactly the same productions. What is needed seems to be more precise non-terminals. One possibility would be to rethink the linguistics behind the Penn Treebank, and ask the annotators to try again. But the original annotation effort took five years, and there is a little appetite for another annotation effort of this scope. Researchers have therefore turned to automated techniques.

### 9.5.1 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 NP $\rightarrow$ NP PP production is more likely if the entire constituent is the child of a VP than if it is the child of S. The observed statistics are (Johnson, 1998):

$$P(\text{NP} \rightarrow \text{NP PP}) = 11\% \quad [9.28]$$

$$P(\text{NP UNDER S} \rightarrow \text{NP PP}) = 9\% \quad [9.29]$$

$$P(\text{NP UNDER VP} \rightarrow \text{NP PP}) = 23\%. \quad [9.30]$$

This phenomenon can be captured by **parent annotation** (Johnson, 1998), in which each non-terminal is augmented with the identity of its parent, as shown in Figure 9.4). This is sometimes called **vertical Markovization**, since a Markov dependency is introduced between each node and its parent (Klein and Manning, 2003). It is analogous to moving from a bigram to a trigram context in a hidden Markov model. In principle, parent annotation squares the size of the set of non-terminals, which could make parsing considerably less efficient. But in practice, the increase in the number of non-terminals that actually appear in the data is relatively modest (Johnson, 1998).

Parent annotation weakens the WCFG independence assumptions. This improves accuracy by enabling the parser to make more fine-grained distinctions, which better capture real linguistic phenomena. However, each production is more rare, and so careful smoothing or regularization is required to control the variance over production scores.

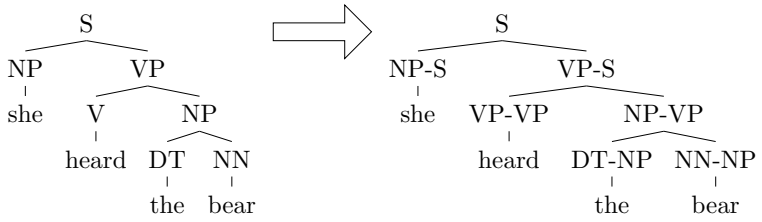



Figure 9.4: Parent annotation in a CFG derivation

### 9.5.2 Lexicalized context-free grammars

The examples in § 9.2.2 demonstrate the importance of individual words in resolving parsing ambiguity: the preposition *on* is more likely to attachment to *met*, while the preposition *of* is more likely to attachment to *President*. But of all word pairs, which are relevant to attachment decisions? Consider the following variants on the original examples:

(9.4) We met the President of Mexico.

(9.5) We met the first female President of Mexico.

(9.6) They had supposedly met the President on Monday.

The underlined words are the **head words** of their respective phrases: *met* and *meet* are the heads of the verb phrase, and *President* is the head of the direct object noun phrase. These heads provide useful 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,

$$\text{NP}(\textit{President}) \rightarrow \text{NP}(\textit{President})\ \text{PP}(\textit{of}) \quad [9.31]$$

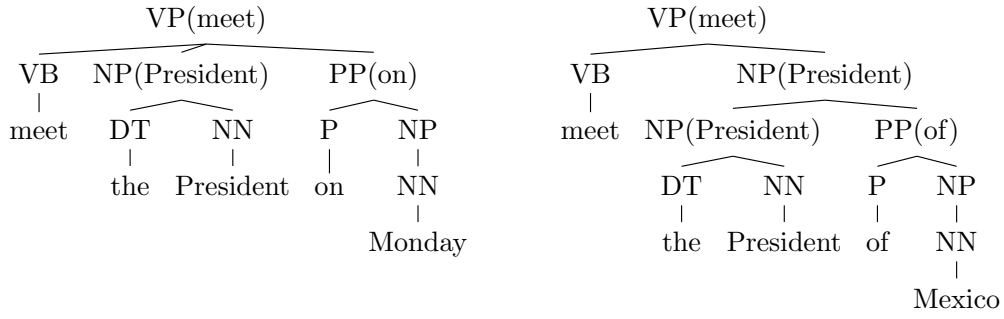
$$\text{NP}(\textit{President}) \rightarrow \text{NP}(\textit{President})\ \text{PP}(\textit{on}). \quad [9.32]$$

Lexicalization was a major step towards accurate PCFG parsing. 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.

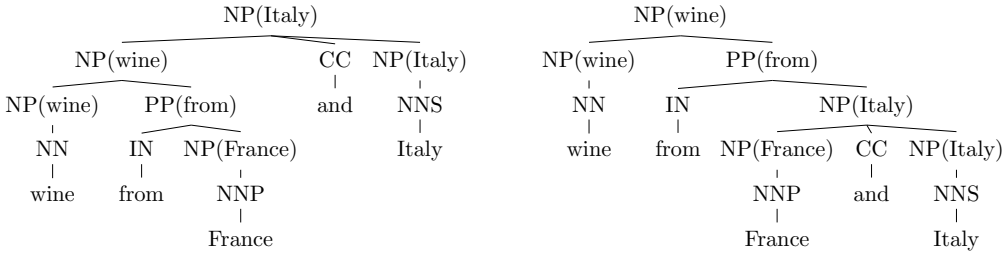
#### Identifying head words

The head of a constituent is the word that is the most useful for determining how that constituent is integrated into the rest of the sentence.[3] The head word of a constituent is

[3]This is a pragmatic definition, befitting our goal of using head words to improve parsing; for a more formal definition, see (Bender, 2013, chapter 7).




(a) Lexicalization and attachment ambiguity




(b) Lexicalization and coordination scope ambiguity

Figure 9.5: Examples of lexicalization

determined recursively: for any non-terminal production, the head of the left-hand side must be the head of one of the children. The head is typically selected according to a set of deterministic rules, sometimes called **head percolation rules**. In many cases, these rules are straightforward: the head of a noun phrase in a NP $\rightarrow$ DET NN production is the head of the noun; the head of a sentence in a S $\rightarrow$ NP VP production is the head of the verb phrase.

Table 9.3 shows a fragment of the head percolation rules used in many English parsing systems. The meaning of the first rule is that to find the head of an S constituent, first look for the rightmost VP child; if you don't find one, then look for the rightmost SBAR child, and so on down the list. Verb phrases are headed by left verbs (the head of *can plan on walking* is *planned*, since the modal verb *can* is tagged MD); noun phrases are headed by the rightmost noun-like non-terminal (so the head of *the red cat* is *cat*),[4] and prepositional phrases are headed by the preposition (the head of *at Georgia Tech* is *at*). Some of these rules are somewhat arbitrary — there's no particular reason why the head of *cats and dogs*

[4]The 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.

| 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 9.3: A fragment of head percolation rules for English, from `http://www.cs.columbia.edu/~mcollins/papers/heads`

should be *dogs* — but the point here is just to get some lexical information that can support parsing, not to make deep claims about syntax. Figure 9.5 shows the application of these rules to two of the running examples.

**Parsing lexicalized context-free grammars**

A naïve application of lexicalization would simply increase the set of non-terminals by taking the cross-product with the set of terminal symbols, so that the non-terminals now include symbols like NP(*President*) and VP(*meet*). Under this approach, the CKY parsing algorithm could be applied directly to the lexicalized production rules. However, the complexity would be cubic in the size of the vocabulary of terminal symbols, which would 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 non-terminal $X$ spanning $\boldsymbol{w}_{i+1:j}$, with head word $h$, where $i < h \leq j$. To compute such a table recursively, we must consider the possibility that each cell 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 $t_\ell$ and $t_r$ respectively, leading to the following recurrence:

$$t_\ell[i, j, h, X] = \max_{(X \rightarrow YZ)} \max_{k>h} \max_{k<h' \leq j} t[i, k, h, Y] \otimes t[k, j, h', Z] \otimes \psi(X(h) \rightarrow Y(h)Z(h')) \quad [9.33]$$

$$t_r[i, j, h, X] = \max_{(X \rightarrow YZ)} \max_{k<h} \max_{i<h' \leq k} t[i, k, h', Y] \otimes t[k, j, h, Z] \otimes (\psi(X(h) \rightarrow Y(h')Z(h))) \quad [9.34]$$

$$t[i, j, h, X] = \max\left(t_\ell[i, j, h, X], t_r[i, j, h, X]\right). \quad [9.35]$$

To compute $t_\ell$, we maximize over all split points $k > h$, since the head word must be in the left child. We then maximize again over possible head words $h'$ for the right child. An analogous computation is performed for $t_r$. The size of the table is now $\mathcal{O}(M^3N)$, where $M$ is the length of the input and $N$ is the number of non-terminals. Furthermore, each

cell is computed by performing $\mathcal{O}(M^2)$ operations, since we maximize over both the split point $k$ and the head $h'$. The time complexity of the algorithm is therefore $\mathcal{O}(RM^5N)$, where $R$ is the number of rules in the grammar. Fortunately, more efficient solutions are possible. In general, the complexity of parsing can be reduced to $\mathcal{O}(M^4)$ in the length of the input; for a broad class of lexicalized CFGs, the complexity can be made cubic in the length of the input, just as in unlexicalized CFGs (Eisner, 2000).

**Estimating lexicalized context-free grammars**

The final problem for lexicalized parsing is how to estimate weights for lexicalized productions $X(i) \rightarrow Y(j)\ Z(k)$. These productions are said to be **bilexical**, because they involve scores over pairs of words: in the example *meet the President of Mexico*, we hope to choose the correct attachment point by modeling the bilexical affinities of (*meet*, *of*) and (*President*, *of*). The number of such word pairs is quadratic in the size of the vocabulary, making it difficult to estimate the weights of lexicalized production rules directly from data. This is especially true for probabilistic context-free grammars, in which the weights are obtained from smoothed relative frequency. In a treebank with a million tokens, a vanishingly small fraction of the possible lexicalized productions will be observed more than once.[5] The Charniak (1997) and Collins (1997) parsers therefore focus on approximating the probabilities of lexicalized productions, using various smoothing techniques and independence assumptions.

In discriminatively-trained weighted context-free grammars, the weights 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 NP(*President*) $\rightarrow$ NP(*President*) PP(*of*) can be computed from the following features:

$$\begin{aligned}\boldsymbol{f}(\text{NP}(\textit{President}) \rightarrow \text{NP}(\textit{President})\ \text{PP}(\textit{of})) = \{&\text{NP}(*) \rightarrow \text{NP}(*)\ \text{PP}(*),\\ &\text{NP}(\textit{President}) \rightarrow \text{NP}(\textit{President})\ \text{PP}(*),\\ &\text{NP}(*) \rightarrow \text{NP}(*)\ \text{PP}(\textit{of}),\\ &\text{NP}(\textit{President}) \rightarrow \text{NP}(\textit{President})\ \text{PP}(\textit{of})\}\end{aligned}$$

The first feature scores the unlexicalized production NP $\rightarrow$ NP PP; the next two features lexicalize only one element of the production, thereby scoring the appropriateness of NP attachment for the individual words *President* and *of*; the final feature scores the specific bilexical affinity of *President* and *of*. For bilexical pairs that are encountered frequently in the treebank, this bilexical feature can play an important role in parsing; for pairs that are absent or rare, regularization will drive its weight to zero, forcing the parser to rely on the more coarse-grained features.

[5] The 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).

In chapter 13, we will encounter techniques for clustering words based on their **distributional** properties — the contexts in which they appear. Such a clustering would group rare and common words, such as *whale, shark, beluga, Leviathan*. Word clusters can be used as features in discriminative lexicalized parsing, striking a middle ground between full lexicalization and non-terminals (Finkel et al., 2008). In this way, labeled examples containing relatively common words like *whale* can help to improve parsing for rare words like *beluga*, as long as those two words are clustered together.

### 9.5.3 *Refinement grammars

Lexicalization improves on context-free parsing by adding detailed information in the form of lexical heads. However, estimating the scores of lexicalized productions is difficult. Klein and Manning (2003) argue that the right level of linguistic detail is somewhere between treebank categories and individual words. Some parts-of-speech and non-terminals are truly substitutable: for example, *cat*/N and *dog*/N. But others are not: for example, the preposition *of* exclusively attaches to nouns, while the preposition *as* is more 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 § 4.2.3. A latent variable PCFG can be estimated by expectation-maximization (Matsuzaki et al., 2005):

- 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 § 9.4.3. In the special case of a PCFG, the inside and outside recur-

rences have the following probabilistic interpretations:

$$\alpha(i,j,X) = \sum_{(X \rightarrow Y\ Z)} \sum_{k=i+1}^{j} \psi(X \rightarrow Y\ Z) \times \alpha(i,k,Y) \times \alpha(k,j,Z) \quad [9.36]$$

$$= \text{p}(\boldsymbol{w}_{i+1:j} \mid X) \quad [9.37]$$

$$\beta(i,j,X) = \sum_{(Y \rightarrow Z\ X)} \sum_{k=0}^{i} \psi(Y \rightarrow Z\ X) \times \alpha(k,i,Z) \times \beta(k,j,Y)$$

$$+ \sum_{(Y \rightarrow X\ Z)} \sum_{k=j+1}^{M} \psi(Y \rightarrow X\ Z) \times \alpha(j,k,Z) \times \beta(i,k,Y) \quad [9.38]$$

$$= \text{p}(\boldsymbol{w}_{1:i}, \boldsymbol{w}_{j+1:M}, X). \quad [9.39]$$

From the inside and outside variables, it is possible to compute the marginal probabilities of anchored productions:

$$\text{p}(X \rightarrow Y\ Z, (i,j,k) \mid \boldsymbol{w}) = \frac{\beta(i,j,X) \times \psi(X \rightarrow Y\ Z) \times \alpha(i,k,Y) \times \alpha(k,j,Z)}{\alpha(0,M,\text{S})}. \quad [9.40]$$

- In the M-step, recompute the parameters of the grammar, by summing over the probabilities of anchored productions,

$$E[\text{count}(X \rightarrow Y\ Z)] = \sum_{i,j,k} \text{p}(X \rightarrow Y\ Z, (i,j,k) \mid \boldsymbol{w}). \quad [9.41]$$

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 9.4. 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.

## 9.6 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-terminals, via parent-annotation, lexicalization, and automated refinement. However, the

| | | | |
|---|---|---|---|
| Proper nouns | | | |
| NNP-14 | *Oct.* | *Nov.* | *Sept.* |
| NNP-12 | *John* | *Robert* | *James* |
| NNP-2 | *J.* | *E.* | *L.* |
| NNP-1 | *Bush* | *Noriega* | *Peters* |
| NNP-15 | *New* | *San* | *Wall* |
| NNP-3 | *York* | *Francisco* | *Street* |
| Personal Pronouns | | | |
| PRP-0 | *It* | *He* | *I* |
| PRP-1 | *it* | *he* | *they* |
| PRP-2 | *it* | *them* | *him* |

Table 9.4: Examples of automatically refined non-terminals and some of the words that they generate (Petrov et al., 2006).

inherent limitations to the expressiveness of context-free parsing motivate the consideration of other search strategies. These strategies abandon the optimality guaranteed by bottom-up parsing, in exchange for the freedom to consider arbitrary properties of the proposed parses.

### 9.6.1 Reranking

A simple way to relax the restrictions of context-free parsing is to perform a two-stage process, in which a context-free parser generates a $k$-best list of candidates, and a **reranker** then selects the best parse from this list (Charniak and Johnson, 2005; Collins and Koo, 2005). The reranker can be trained from an objective that is similar to multi-class classification: the goal is to learn weights that assign a high score to the reference parse, or to the parse on the $k$-best list that has the lowest error. In either case, the reranker need only evaluate the $K$ best parses, and so no context-free assumptions are necessary. This opens the door to more expressive scoring functions:

- It is possible to incorporate arbitrary non-local features, such as the structural parallelism and right-branching orientation of the parse (Charniak and Johnson, 2005).
- Reranking enables the use of **recursive neural networks**, in which each constituent span $\boldsymbol{w}_{i:j}$ receives a vector $\boldsymbol{u}_{i:j}$ which is computed from the vector representations of its children, using a composition function that is linked to the production rule (Socher et al., 2013), e.g.,

$$\boldsymbol{u}_{i:j} = f\left(\Theta_{X \to Y\ Z} \begin{bmatrix} \boldsymbol{u}_{i:k-1} \\ \boldsymbol{u}_{k:j} \end{bmatrix}\right) \quad [9.42]$$

The overall score of the parse can then be computed from the final vector, $\Psi(\tau) = \boldsymbol{\theta}\boldsymbol{u}_{1:M}$.

Reranking can yield substantial improvements in accuracy. The main limitation is that it can only find the best parse among the $K$-best offered by the generator, so it is inherently limited by the ability of the bottom-up parser to find high-quality candidates.

### 9.6.2 Transition-based parsing

Structure prediction can be viewed as a form of search. An alternative search strategy is to read the input from left-to-right, gradually building up a parse structure through a series of **transitions**. Transition-based parsing is described in more detail in the next chapter, in the context of dependency parsing. However, it can also be applied to CFG parsing, as briefly described here.

For any context-free grammar, there is an equivalent **pushdown automaton**, a model of computation that accepts exactly those strings that can be derived from the grammar. This computational model consumes the input from left to right, while pushing and popping elements on a stack. This architecture provides a natural transition-based parsing framework for context-free grammars, known as **shift-reduce parsing**.

Shift-reduce parsing is a type of transition-based parsing, in which the parser can take the following actions:

- *shift* the next terminal symbol onto the stack;
- *unary-reduce* the top item on the stack, using a unary production rule in the grammar;
- *binary-reduce* the top two items onto the stack, using a binary production rule in the grammar.

The set of available actions is constrained by the situation: the parser can only shift if there are remaining terminal symbols in the input, and it can only reduce if an applicable production rule exists in the grammar. If the parser arrives at a state where the input has been completely consumed, and the stack contains only the element S, then the input is accepted. If the parser arrives at a non-accepting state where there are no possible actions, the input is rejected. A parse error occurs if there is some action sequence that would accept an input, but the parser does not find it.

**Example** Consider the input *we eat sushi* and the grammar in Table 9.1. The input can be parsed through the following sequence of actions:

1. **Shift** the first token *we* onto the stack.

2. **Reduce** the top item on the stack to NP, using the production NP $\rightarrow$ *we*.
3. **Shift** the next token *eat* onto the stack, and **reduce** it to V with the production V $\rightarrow$ *eat*.
4. **Shift** the final token *sushi* onto the stack, and **reduce** it to NP. The input has been completely consumed, and the stack contains [NP, V, NP].
5. **Reduce** the top two items using the production VP $\rightarrow$ V NP. The stack now contains [VP, NP].
6. **Reduce** the top two items using the production S $\rightarrow$ NP VP. The stack now contains [S]. Since the input is empty, this is an accepting state.

One thing to notice from this example is that the number of shift actions is equal to the length of the input. The number of reduce actions is equal to the number of non-terminals in the analysis, which grows linearly in the length of the input. Thus, the overall time complexity of shift-reduce parsing is linear in the length of the input (assuming the complexity of each individual classification decision is constant in the length of the input). This is far better than the cubic time complexity required by CKY parsing.

**Transition-based parsing as inference** In general, it is not possible to guarantee that a transition-based parser will find the optimal parse, $\text{argmax}_\tau \Psi(\tau; \boldsymbol{w})$, even under the usual CFG independence assumptions. We could assign a score to each anchored parsing action in each context, with $\psi(a, c)$ indicating the score of performing action $a$ in context $c$. One might imagine that transition-based parsing could efficiently find the derivation that maximizes the sum of such scores. But this too would require backtracking and searching over an exponentially large number of possible action sequences: if a bad decision is made at the beginning of the derivation, then it may be impossible to recover the optimal action sequence without backtracking to that early mistake. This is known as a **search error**. Transition-based parser can incorporate arbitrary features, without the restrictive independence assumptions required by chart parsing; search errors are the price that must be paid for this flexibility.

**Learning transition-based parsing** Transition-based parsing can be combined with machine learning by training a classifier to select the correct action at each position. This classifier is free to choose any feature of the input, the state of the parser, and the parse history. However, there is no optimality guarantee: the parser may choose a suboptimal parse, due to a mistake at the beginning of the analysis. Nonetheless, some of the strongest CFG parsers are based on the shift-reduce architecture, rather than CKY. A recent generation of models links shift-reduce parsing with recurrent neural networks, updating a hidden state vector while consuming the input (e.g., Cross and Huang, 2016; Dyer et al., 2016). Learning algorithms for transition-based parsing are discussed in more detail in § 10.3.

## Exercises

1. Consider the following PCFG:

$$\text{p}(X \rightarrow X\ X) = \frac{1}{2} \quad [9.43]$$

$$\text{p}(X \rightarrow Y) = \frac{1}{2} \quad [9.44]$$

$$\text{p}(Y \rightarrow \sigma) = \frac{1}{|\Sigma|}, \forall \sigma \in \Sigma \quad [9.45]$$

   a) Compute the probability $\text{p}(\hat{\tau})$ of the maximum probability parse for a string $\boldsymbol{w} \in \Sigma^M$.

   b) Compute the marginal probability $\text{p}(\boldsymbol{w}) = \sum_{\tau:\text{yield}(\tau)=\boldsymbol{w}} \text{p}(\tau)$.

   c) Compute the conditional probability $\text{p}(\hat{\tau} \mid \boldsymbol{w})$.

2. Use the inside and outside scores to compute the marginal probability $\text{p}(X_{i:j} \rightarrow Y_{i:k-1}\ Z_{k:j} \mid \boldsymbol{w})$, indicating that $Y$ spans $\boldsymbol{w}_{i:k-1}$, $Z$ spans $\boldsymbol{w}_{k:j}$, and $X$ is the parent of $Y$ and $Z$, spanning $\boldsymbol{w}_{i:j}$.

3. Suppose that the potentials $\Psi(X \rightarrow \alpha)$ are log-probabilities, so that $\sum_\alpha \exp \Psi(X \rightarrow \alpha) = 1$ for all $X$. Verify that the semiring inside recurrence from Equation 9.21 generates the log-probability $\log \text{p}(\boldsymbol{w}) = \log \sum_{\tau:\text{yield}(\tau)=\boldsymbol{w}} \text{p}(\tau)$.

4. more exercises tk

# Chapter 10

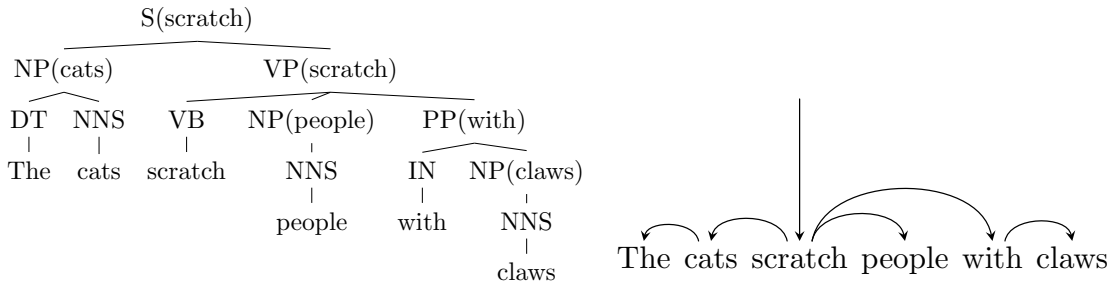
# Dependency Parsing

The previous chapter discussed algorithms for analyzing sentences in terms of nested constituents, such as noun phrases and verb phrases. However, many of the key sources of ambiguity in phrase-structure analysis relate to questions of **attachment**: where to attach a prepositional phrase or complement clause, how to scope a coordinating conjunction, and so on. These attachment decisions can be represented with a more lightweight structure: a directed graph over the words in the sentence, known as a **dependency parse**. In recent years, syntactic annotation has shifted its focus to such dependency structures: at the time of this writing, the **Universal Dependencies** project offers more than 100 dependency treebanks for more than 60 languages.[1] This chapter will describe the linguistic ideas underlying dependency grammar, and then discuss exact and transition-based parsing algorithms. The chapter will also discuss recent research on **learning to search** in transition-based structure prediction.

## 10.1 Dependency grammar

While **dependency grammar** has a rich history of its own (Tesnière, 1966; Kübler et al., 2009), it can be motivated by extension from the lexicalized context-free grammars that we encountered in previous chapter (§ 9.5.2). Recall that lexicalization augments each non-terminal with a **head word**. The head of a constituent is identified recursively, using a set of **head rules**, as shown in Table 9.3. An example of a lexicalized context-free parse is shown in Figure 10.1a. In this sentence, the head of the S constituent is the main verb, *scratch*; this non-terminal then produces the noun phrase *the cats*, whose head word is *cats*, and from which we finally derive the word *the*. Thus, the word *scratch* occupies the central position for the sentence, with the word *cats* playing a supporting role. In turn, *cats*

[1] `universaldependencies.org`




(a) lexicalized constituency parse (b) unlabeled dependency tree

Figure 10.1: Dependency grammar is closely linked to lexicalized context free grammars: each lexical head has a dependency path to every other word in the constituent.

occupies the central position for the noun phrase, with the word *the* playing a supporting role.

These relationships, which hold between the words in the sentence, can be formalized in a directed graph, by placing an edge from word $i$ to word $j$ iff word $i$ is the head of the first branching node above a node headed by $j$. Thus, in our example, we would have *scratch* $\rightarrow$ *cats* and *cats* $\rightarrow$ *the*. We would not have the edge *scratch* $\rightarrow$ *the*, because although S(*scratch*) dominates DET(*the*) in the graph, it is not the *first* branching node above the determiner. These edges describe **syntactic dependencies**, a bilexical relationship between a **head** and a **dependent**, which is at the heart of dependency grammar.

If we continue to build out this **dependency graph**, we will eventually reach every word in the sentence, as shown in Figure 10.1b. In this graph — and in all graphs constructed in this way — every word has exactly one incoming edge, except for the root word, which is indicated by a special incoming arrow from above. Furthermore, the graph is *weakly connected*: if the directed edges were replaced with undirected edges, there would be a path between all pairs of nodes. From these properties, it can be shown that there are no cycles in the graph (or else at least one node would have to have more than one incoming edge), and therefore, the graph is a tree.

### 10.1.1 Heads and dependents

A dependency edge implies an asymmetric syntactic relationship between the head and dependent words, sometimes called **modifiers**. For a pair like *the cats* or *cats scratch*, how do we decide which is the head? Here are some possible criteria:

- The head sets the syntactic category of the construction: for example, nouns are the heads of noun phrases, and verbs are the heads of verb phrases.

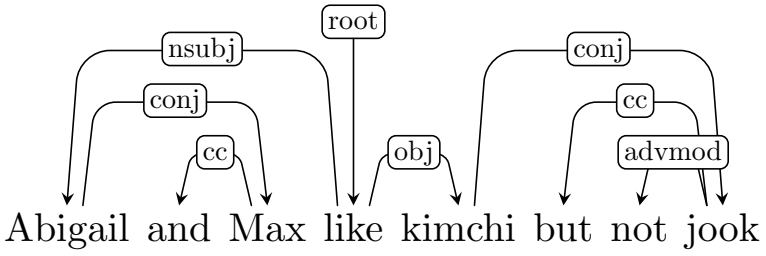



Figure 10.2: In the Universal Dependencies annotation system, the left-most item of a coordination is the head.

- The modifier may be optional while the head is mandatory: for example, in the sentence *cats scratch people with claws*, the subtrees *cats scratch* and *cats scratch people* are grammatical sentences, but *with claws* is not.
- The head determines the morphological form of the modifier: for example, in languages that require gender agreement, the gender of the noun determines the gender of the adjectives and determiners.
- Edges should first connect content words, and then connect function words.

As always, these guidelines sometimes conflict. The Universal Dependencies (UD) project has attempted to identify a set of principles that can be applied to dozens of different languages (Nivre et al., 2016).[2] These guidelines are based on the universal part-of-speech tags from chapter 7. They differ somewhat from the head rules described in § 9.5.2: for example, on the principle that dependencies should relate content words, the prepositional phrase *with claws* would be headed by *claws*, resulting in an edge *scratch* $\rightarrow$ *claws*, and another edge *claws* $\rightarrow$ *with*.

One objection to dependency grammar is that not all syntactic relations are asymmetric. Coordination is one of the most obvious examples (Popel et al., 2013): in the sentence, *Abigail and Max like kimchi* (Figure 10.2), which word is the head of the coordinated noun phrase *Abigail and Max*? Choosing either *Abigail* or *Max* seems arbitrary; fairness argues for the choice of *and*, but this seems the least important word in the noun phrase, and selecting it would violate the principle of linking content words first. The Universal Dependencies annotation system arbitrarily chooses the left-most item as the head — in this case, *Abigail* — and includes edges from this head to both *Max* and the coordinating conjunction *and*. These edges are distinguished by the labels CONJ (for the thing begin conjoined) and CC (for the coordinating conjunction). The labeling system is discussed next.

[2]The latest and most specific guidelines are available at `universaldependencies.org/guidelines.html`

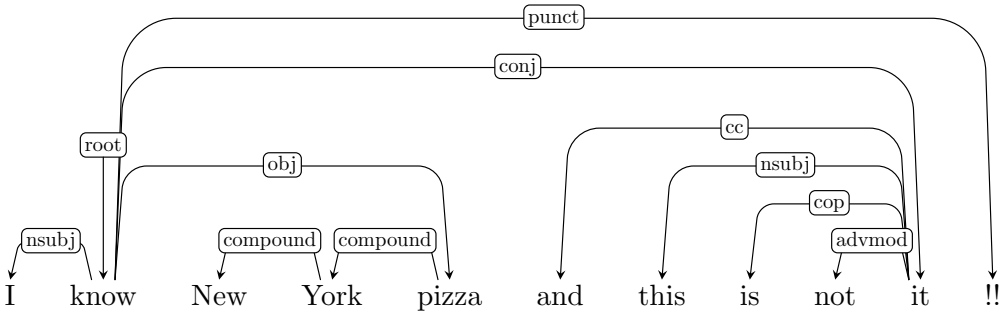



Figure 10.3: A labeled dependency parse from the UD Treebank (reviews-361348-0006)

### 10.1.2 Labeled dependencies

Edges may be **labeled** to indicate the nature of the syntactic relation that holds between the two elements. For example, in Figure 10.2, the label NSUBJ on the edge from *like* to *Abigail* indicates that the subtree headed by *Abigail* is the noun subject of the predicate verb *like*; similarly, the label OBJ on the edge from *like* to *kimchi* indicates that the subtree headed by *kimchi* is the object.[3] The negation *not* is treated as an adverbial modifier (ADVMOD) on the noun *jook*.

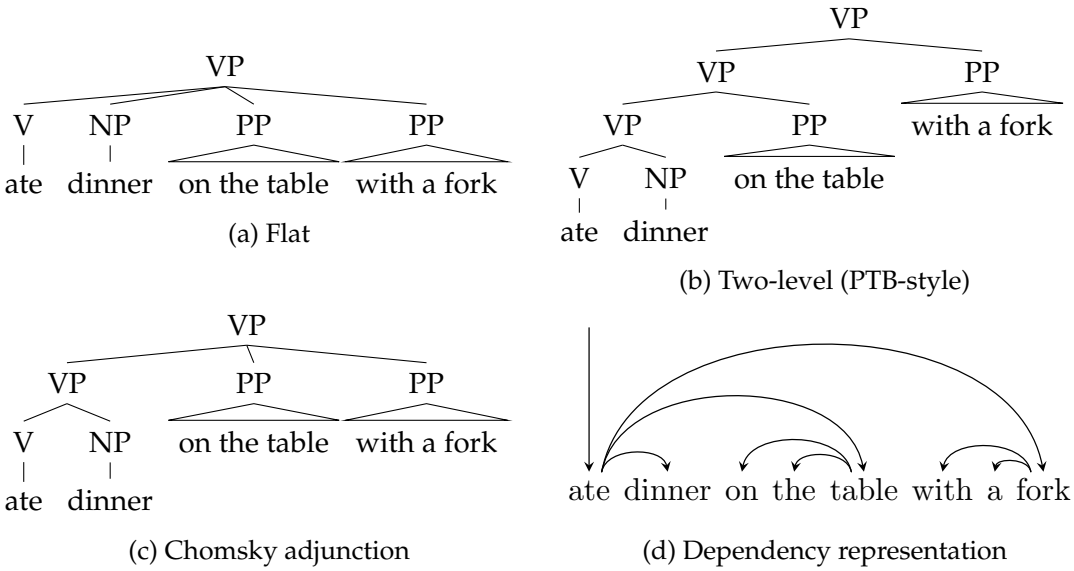
A slightly more complex example is shown in Figure 10.3. The multiword expression *New York Pizza* is treated as a "flat" unit of text, with the elements linked by the COMPOUND relation. The sentence includes two clauses that are conjoined in the same way that noun phrases are conjoined in Figure 10.2. The second clause contains a **copula** verb (see § 7.1.1). For such clauses, we treat the "object" of the verb as the root — in this case, *it* — and label the verb as a dependent, with the COP relation. This example also shows how punctuations are treated, with label PUNCT.

### 10.1.3 Dependency subtrees and constituents

Dependency trees hide information that would be present in a CFG parse. Often what is hidden is in fact irrelevant: for example, Figure 10.4 shows three different ways of 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 10.4d, these three cases all look the same in a dependency parse, which is an advantage of the dependency representation.

But dependency grammar imposes its own set of annotation decisions, such as the

[3]Earlier 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.




(a) Flat

(b) Two-level (PTB-style)

(c) Chomsky adjunction

(d) Dependency representation

Figure 10.4: The three different CFG analyses of this verb phrase all correspond to a single dependency structure.

identification of the head of a coordination (§ 10.1.1); (unlexicalized) 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 node. 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 is 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:

**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.*

Figure 10.5 gives an example of a non-projective dependency graph in English. This dependency graph does not correspond to any constituent parse. In languages where

| | % non-projective edges | % non-projective sentences |
|---|---|---|
| Czech | 1.86% | 22.42% |
| English | 0.39% | 7.63% |
| German | 2.33% | 28.19% |

Table 10.1: Frequency of non-projective dependencies in three languages (Kuhlmann and Nivre, 2010)

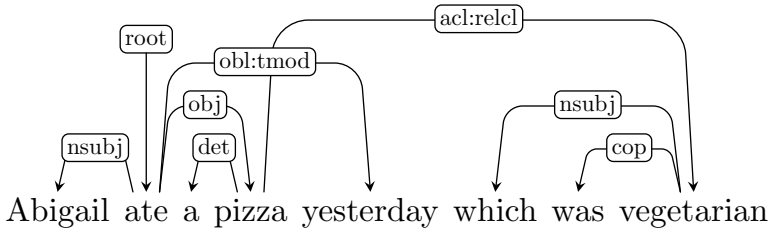



Figure 10.5: An example of a non-projective dependency parse. The "crossing edge" arises from the relative clause *which was vegetarian* and the oblique temporal modifier *yesterday*.

non-projectivity is common, such as Czech and German, it is better to annotate dependency trees directly, rather than deriving them from constituent parses. An example is the Prague Dependency Treebank (Böhmová et al., 2003), which contains 1.5 million words of Czech, with approximately 12,000 non-projective edges (see Table 10.1). Even though relatively few dependencies are non-projective in Czech and German, many sentences have at least one such dependency. As we will soon see, projectivity has important algorithmic consequences.

## 10.2 Graph-based dependency parsing

Let $\boldsymbol{y} = \{\langle i, j, r\rangle\}$ indicate a dependency graph with relation $r$ from head word $w_i$ to modifier $w_j$. Given a scoring function $\Psi(\boldsymbol{y}, \boldsymbol{w}; \boldsymbol{\theta})$, the optimal parse is,

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{w})}{\text{argmax}}\, \Psi(\boldsymbol{y}, \boldsymbol{w}; \boldsymbol{\theta}). \quad [10.1]$$

The learning problem is to minimize $\boldsymbol{\theta}$ with respect to a loss on a labeled dataset, $\left(\boldsymbol{y}^{(1:N)}, \boldsymbol{w}^{(1:N)}\right)$. As usual, the number of possible labelings $\mathcal{Y}(\boldsymbol{w})$ is exponential in the length of the input. In the case of non-projective dependency parsing, the set $\mathcal{Y}(\boldsymbol{w})$ includes all possible spanning trees over a complete graph with $M$ nodes, where $M$ is the length of the sentence $\boldsymbol{w}$. The size of this set is $M^{M-2}$ (Wu and Chao, 2004). Algorithms that search over this space of possible graphs are known as **graph-based dependency 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 $i \rightarrow j$:

$$\Psi(\boldsymbol{y}, \boldsymbol{w}; \boldsymbol{\theta}) = \bigotimes_{\langle i,j,r \rangle \in \boldsymbol{y}} \psi(\boldsymbol{w}, i, j, r; \boldsymbol{\theta}). \quad [10.2]$$

Dependency parsers that operate under this assumption are known as **arc-factored**, since the overall score is a product of scores over all arcs. If each score $\psi$ is a probability, the the $\otimes$ operator is multiplication, and the score for the entire parse is also a probability (Eisner, 1996). If each score is a log-probability or is the output of a discriminatively-trained scoring function, the $\otimes$ operator is addition.

**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 [todo: figure].

$$\Psi(\boldsymbol{y}, \boldsymbol{w}; \boldsymbol{\theta}) = \bigotimes_{\langle i,j,r \rangle \in \boldsymbol{y}} \bigotimes_{k, r' : \langle k,i,r' \rangle \in \boldsymbol{y}} \psi(\boldsymbol{w}, i, j, k, r, r'; \boldsymbol{\theta}) \\ \bigotimes_{s, r' : \langle i,s,r' \rangle \in \boldsymbol{y} \wedge s \neq j} \psi(\boldsymbol{w}, i, j, s, r, r'; \boldsymbol{\theta}). \quad [10.3]$$

The top line scores computes a scoring function that includes the grandparent $k$; the bottom line computes a scoring function for each sibling $s$. For projective dependency graphs, there are efficient algorithms for second-order and third-order dependency parsing (Eisner, 1996; McDonald and Pereira, 2006; Koo and Collins, 2010); for non-projective dependency graphs, second-order dependency parsing is NP-hard (McDonald and Pereira, 2006). The specific algorithms are discussed in § 10.2.1.

### 10.2.1 Graph-based parsing algorithms

The distinction between projective and non-projective dependency trees (§ 10.1.3) plays a key role in the choice of algorithms. Because projective dependency trees are closely related to (and can be derived from) lexicalized constituent trees, lexicalized parsing algorithms can be applied directly. For the more general problem of parsing to arbitrary spanning trees, a different class of algorithms is required. In both cases, arc-factored dependency parsing relies on precomputing the scores $\psi(i, j, r)$ for each potential edge from $i$ to $j$ with label $r$. There are $\mathcal{O}(M^2 R)$ such scores, where $M$ is the length of the input and $R$ is the number of dependency relation types, and this is a lower bound on the time and space complexity of any exact algorithm for arc-factored dependency parsing.

### Projective dependency parsing

As discussed in § 10.1, any lexicalized constituency tree can be converted into a projective dependency tree by creating arcs between the heads of constituents and their parents. As a result, any algorithm for lexicalized constituent parsing can be converted into an algorithm for projective dependency parsing, by converting arc scores into scores for lexicalized productions. [todo: use this as the basis for an exercise?] As noted in § 9.5.2, there are bottom-up cubic time algorithms for lexicalized constituent parsing, which are extensions of the CKY algorithm. Therefore, arc-factored projective dependency parsing can be performed in cubic time in the length of the input.

Second-order projective dependency parsing can also be performed in cubic time, with minimal modifications to the lexicalized parsing algorithm (Eisner, 1996). It is possible to go even further, to **third-order dependency parsing**, in which the scoring function may consider great-grandparents, grand-siblings, and "tri-siblings", as shown in **??**. Third-order dependency parsing can be performed in $\mathcal{O}(M^4)$ time, which can be made practical through the use of pruning to eliminate unlikely edges (Koo and Collins, 2010).

### Non-projective dependency parsing

In non-projective dependency parsing, the goal is to identify the highest-scoring spanning tree over the words in the sentence. The arc-factored assumption ensures that the score for each spanning tree will be computed as a sum over scores for the edges, $\psi(i, j, r)$, which are precomputed. Based on these scores, we build a weighted connected graph. Arc-factored non-projective dependency parsing is then equivalent to finding the the spanning tree that achieves the maximum total score, $\Psi(\boldsymbol{y}; \boldsymbol{w}) = \bigotimes_{\langle i,j,r\rangle \in \boldsymbol{y}} \psi(i, j, r)$. The **Chu-Liu-Edmonds algorithm** (Chu and Liu, 1965; Edmonds, 1967) computes this spanning tree in time $\mathcal{O}(M^3 R)$. The algorithm, which is sketched in Algorithm 14, operates recursively. It first identifies the highest scoring incoming edge for each node, and then checks the graph for cycles. If there are no cycles, the resulting graph is a spanning tree, and moreover, it is the maximum spanning tree, because there is no better-scoring incoming edge for each node. If there is a cycle, the cycle is collapsed into a "super-node", whose incoming edges have scores equal to the scores of the best spanning tree that includes both the edge and all nodes in the cycle. The algorithm works because it can be proved that the maximum spanning tree on the contracted graph is equivalent to the maximum spanning tree on the original graph.

Let us consider the time complexity of unlabeled dependency parsing first. For each of the $M$ words in the sentence, one must search all $M - 1$ other words for the highest-scoring incoming edge, for a time complexity of $\mathcal{O}(M^2)$. In the worst case, it is necessary to contract the graph $M$ times. If we redo the search within each contraction, the total cost would be $\mathcal{O}(M^3)$ in the length of the input. However, further optimizations are possible, reducing the complexity of $\mathcal{O}(M^2)$(Tarjan, 1977). To generalize the algorithm to labeled

**Algorithm 14** Chu-Liu-Edmonds algorithm for unlabeled dependency parsing

```
procedure CHU-LIU-EDMONDS(\{\psi(i \rightarrow j)\}_{i,j \in \{1 \ldots M\}})
    for j \in 1 \ldots M do
        h_j \leftarrow \text{argmax}_i \psi(i \rightarrow j)        ▷ Find the best incoming edge for each node
    \tau \leftarrow \{j, h_j\}_{j \in 1 \ldots M}                  ▷ \tau is the graph of the best incoming edges
    \mathcal{C} \leftarrow FINDCYCLES(\tau)
    if \mathcal{C} = \varnothing then
        return \tau                                            ▷ If \tau has no cycles, it is the best tree
    else                                                       ▷ Otherwise, collapse each cycle
        for each cycle c \in \mathcal{C} do
            Remove all nodes in the cycle from the graph
            Add a "super-node" representing the cycle
        Let G be the resulting graph
        return CHU-LIU-EDMONDS(G)                              ▷ Call recursively on the collapsed graph
```

dependency parsing, it is necessary only to compute the best scoring label for each possible edge. Because of the arc-factoring assumption, the edge labels are decoupled from each other, so this can be done as a preprocessing step, with total complexity of $\mathcal{O}(M^2 R)$.

Given the tractability of higher-order projective dependency parsing, you may be surprised to learn that non-projective second-order dependency parsing is NP-Hard! This can be proved by reduction from the vertex cover problem (Neuhaus and Bröker, 1997). One heuristic solution is to do projective parsing first, and then post-process the projective dependency parse to add non-projective edges (Nivre and Nilsson, 2005). More recent work has applied techniques for approximate inference in graphical models, including belief propagation (Smith and Eisner, 2008), integer linear programming (Martins et al., 2009), variational inference (Martins et al., 2010), and Markov Chain Monte Carlo (Zhang et al., 2014).

### 10.2.2 Computing scores for dependency arcs

The arc-factored scoring function $\psi(\boldsymbol{w}, i, j, r)$ can be defined in several ways:

$$\text{Log-linear} \quad \psi(\boldsymbol{w}, i, j, r; \boldsymbol{\theta}) = \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, i, j, r) \tag{10.4}$$

$$\text{Neural} \quad \psi(\boldsymbol{w}, i, j, r; \boldsymbol{\theta}) = \text{Feedforward}(\boldsymbol{u}_{w_i}, \boldsymbol{u}_{w_j}; \boldsymbol{\theta}) \tag{10.5}$$

$$\text{Generative} \quad \psi(\boldsymbol{w}, i, j, r; \boldsymbol{\theta}) = \text{p}(w_j, r \mid w_i). \tag{10.6}$$

#### Log-linear arc scores

Log-linear models for dependency parsing incorporate many of the same features used in sequence labeling and discriminative constituent parsing. These include:

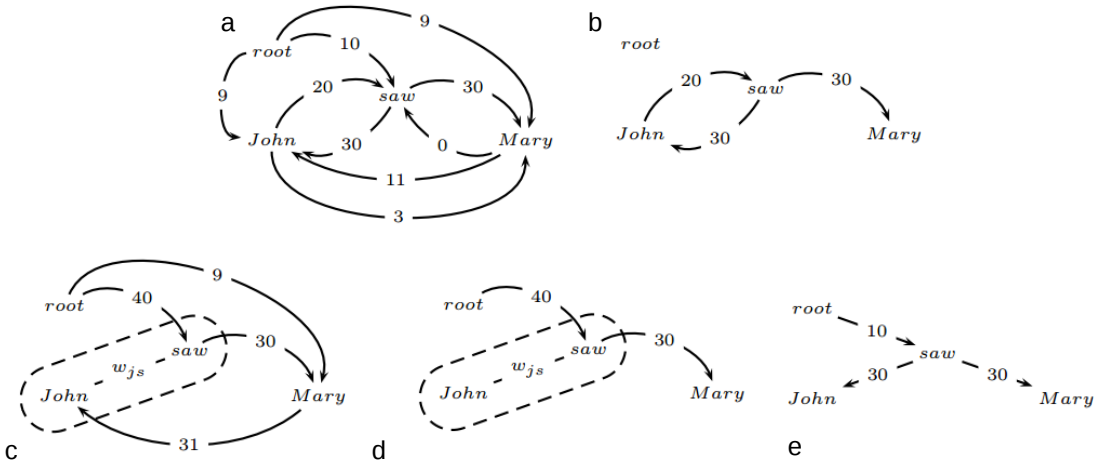



Figure 10.6: An illustration of the MST algorithm on a simple example. Figure borrowed from McDonald et al. (2005).

- the length and direction of the arc;
- the words $w_i$ and $w_j$ linked by the dependency relation;
- the prefixes, suffixes, and part-of-speech of these words;
- the neighbors of the dependency arc, $w_{i-1}, w_{i+1}, w_{j-1}, w_{j+1}$;
- the prefixes, suffixes, and part-of-speech of these neighbor words.

Each of these features can be conjoined with the dependency edge label $r$. Note that features in an arc-factored parser can refer to words other than $w_i$ and $w_j$. The restriction is that the features consider only a single arc.

**Bilexical features** (e.g., *sushi* $\rightarrow$ *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:

$$\begin{aligned}\boldsymbol{f}(\textit{we eat sushi with chopsticks}, 3, 5) = \langle &\textit{sushi} \rightarrow \textit{chopsticks},\\ &\textit{sushi} \rightarrow \textsc{Nns},\\ &\textsc{Nn} \rightarrow \textit{chopsticks},\\ &\textsc{Nns} \rightarrow \textsc{Nn}\rangle.\end{aligned}$$

Regularized discriminative learning algorithms can then learn to trade off between features at varying levels of detail. McDonald et al. (2005) take this approach as far as tetralexical features (e.g., $\langle w_i, w_{i+1}, w_{j-1}, w_j\rangle$). Such features help to avoid choosing arcs that are unlikely due to the intervening words: for example, there is unlikely to be an edge between two nouns if the intervening span contains a verb. A large list of first and second-order features is provided by Bohnet (2010), who uses a hashing function to store these features efficiently.

#### Neural arc scores

Given vector representations $\boldsymbol{v}_i$ for each word $w_i$ in the input, a set of arc scores can be computed from a feedforward neural network:

$$\psi(i, j, r) = \text{FeedForward}(\boldsymbol{v}_i, \boldsymbol{v}_j; \boldsymbol{\theta}_r), \quad [10.7]$$

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,

$$\psi(i, j, r) = \boldsymbol{\beta}_r \tanh(\boldsymbol{\Theta}_r[\boldsymbol{v}_i; \boldsymbol{v}_j] + b_r^{(1)}) + b_r^{(2)}, \quad [10.8]$$

where $\boldsymbol{\Theta}_r$ is a matrix, $\boldsymbol{\beta}_r$ is a vector, and each $b_r$ is a scalar.

The vector $\boldsymbol{v}_i$ can be set equal to the word embedding $\boldsymbol{v}_{w_i}$, which may be pre-trained or learned by backpropagation (Pei et al., 2015). Alternatively, contextual information can be incorporated by applying a bidirectional recurrent neural network across the input,

$$\boldsymbol{v}_i = [\overrightarrow{\boldsymbol{v}}_i; \overleftarrow{\boldsymbol{v}}_i] \quad [10.9]$$

$$\overrightarrow{\boldsymbol{v}}_i = \overrightarrow{\text{RNN}}(\boldsymbol{u})_i \quad [10.10]$$

$$\overleftarrow{\boldsymbol{v}}_i = \overleftarrow{\text{RNN}}(\boldsymbol{u})_i, \quad [10.11]$$

with $\boldsymbol{u}_i$ equal to the embedding of word $w_i$ (Kiperwasser and Goldberg, 2016).

**Probabilistic arc scores**

If each arc score is equal to the probability $\text{p}(w_j, r \mid w_i)$, then the product of scores gives the probability of the sentence and arc labels, by the chain rule. For example, consider the unlabeled parse of *we eat sushi with rice*,

$$\boldsymbol{y} = \{(\text{ROOT}, 2), (2, 1), (2, 3), (3, 5), (5, 4)\} \quad [10.12]$$

$$\text{p}(\boldsymbol{w} \mid \boldsymbol{y}) = \prod_{\langle i,j \rangle \in \boldsymbol{y}} \text{p}(w_j \mid w_i) \quad [10.13]$$

$$= \text{p}(\textit{eat} \mid \text{ROOT}) \times \text{p}(\textit{we} \mid \textit{eat}) \times \text{p}(\textit{sushi} \mid \textit{eat}) \times \text{p}(\textit{rice} \mid \textit{sushi}) \times \text{p}(\textit{with} \mid \textit{rice}). \quad [10.14]$$

Probabilistic generative models are rarely used in supervised dependency parsing, but are used in combination with expectation-maximization for unsupervised dependency parsing (Klein and Manning, 2004).

### 10.2.3 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 $\ell(\boldsymbol{\theta}; \boldsymbol{w}^{(i)}, \boldsymbol{y}^{(i)})$, we can compute gradient-based updates to the parameters. For a model with log-linear arc scores and a perceptron loss, we obtain the usual structured perceptron update,

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{w})}{\text{argmax}}\, \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, \boldsymbol{y}') \quad [10.15]$$

$$\boldsymbol{\theta} = \boldsymbol{\theta} + \boldsymbol{f}(\boldsymbol{w}, \boldsymbol{y}) - \boldsymbol{f}(\boldsymbol{w}, \hat{\boldsymbol{y}}) \quad [10.16]$$

In this case, the $\text{argmax}$ requires a maximization over all dependency trees for the sentence, which can be computed using the algorithms described in § 10.2.1. We can apply all the usual tricks from § 1.2: weight averaging, large-margin, and regularization. McDonald et al. (2005,?) were the first to treat dependency parsing as a structure prediction problem, using MIRA, an online margin-based learning algorithm. Neural arc scores can be learned in the same way, backpropagating from a margin loss to updates on the feed-forward network that computes the score for each edge.

A conditional random field for arc-factored dependency parsing is built on the probability model,

$$\text{p}(\boldsymbol{y} \mid \boldsymbol{w}) = \frac{\exp \sum_{(i,j,r) \in \boldsymbol{y}} \psi(i, j, r)}{\sum_{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{w})} \exp \sum_{(i',j',r') \in \boldsymbol{y}'} \psi(i', j', r')}. \quad [10.17]$$

Such a model is trained to minimize the negative log conditional-likelihood. Just as in CRF sequence models (§ 6.5.3) and the logistic regression classifier (§ 1.4), the gradients involve

marginal probabilities $\text{p}((i, j, r) \mid \boldsymbol{w}; \boldsymbol{\theta})$, which in this case are probabilities over individual edges. For projective dependency trees, the marginal probabilities can be computed in cubic time, using a variant of the inside-outside algorithm (Lari and Young, 1990). For non-projective dependency parsing, marginals can also be computed in cubic time, using the **matrix-tree theorem** (Koo et al., 2007; McDonald et al., 2007; Smith and Smith, 2007). Details of these methods are described by Kübler et al. (2009).

## 10.3 Transition-based dependency parsing

Graph-based dependency parsing offers exact inference, meaning that it is possible to recover the best-scoring parse for any given model. But this comes at a price: the scoring function is required to decompose into local parts — in the case of non-projective parsing, these parts are restricted to individual arcs. These limitations are felt more keenly in dependency parsing than in sequence labeling, because second-order dependency features are critical to correctly identify some types of attachments. For example, prepositional phrase attachment depends on the attachment point, the object of the preposition, and the preposition itself; arc-factored scores cannot account for all three of these features simultaneously. Graph-based dependency parsing may also be criticized on the basis of intuitions about human language processing: people read and listen to sentences *sequentially*, incrementally building mental models of the sentence structure and meaning before getting to the end (Jurafsky, 1996). This seems hard to reconcile with graph-based algorithms, which perform bottom-up operations on the entire sentence, requiring the parser to keep every word in memory. Finally, from a practical perspective, graph-based dependency parsing is relatively slow, running in cubic time in the length of the input.

Transition-based algorithms address all three of these objections. They work by moving through the sentence sequentially, while performing actions that incrementally update a stored representation of what has been read thus far. As in the shift-reduce parser from the previous chapter (§ 9.6.2), 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 analysis 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**.

[4] Transition systems also exist for non-projective dependency parsing (e.g., Nivre, 2008)

### 10.3.1 Transition systems for dependency parsing

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 $\sigma$ is the stack, $\beta$ is the input buffer, and $A$ is the list of arcs that have been created (Nivre, 2008). In the initial configuration,

$$C_{\text{initial}} = ([\text{ROOT}], \boldsymbol{w}, \varnothing), \tag{10.18}$$

indicating that the stack contains only the special root node ROOT, the entire input is on the buffer, and the set of arcs is empty. An accepting configuration is,

$$C_{\text{accept}} = ([\text{ROOT}], \varnothing, A), \tag{10.19}$$

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. It includes the following actions:

- SHIFT: move the first item from the input buffer on to the top of the stack,

$$(\sigma, i|\beta, A) \Rightarrow (\sigma|i, \beta, A), \tag{10.20}$$

  where we write $i|\beta$ to indicate that $i$ is the leftmost item in the input buffer, and $\sigma|i$ to indicate the result of pushing $i$ on to stack $\sigma$.

- ARC-LEFT: create a new left-facing arc between the item on the top of the stack and the first item in the input buffer. This item is then "popped" to the front of the input buffer, and the arc is added to $A$.

$$(\sigma|i, j|\beta, A) \Rightarrow (\sigma, j|\beta, A \oplus (j, i, r)), \tag{10.21}$$

  where $r$ is the label of the dependency arc, and $\oplus$ concatenates the new arc $(j, i, r)$ to the list $A$.

| | $\sigma$ | $\beta$ | action | arc added to $\mathcal{A}$ |
|---|---|---|---|---|
| 1. | [ROOT] | *they like bagels with lox* | SHIFT | |
| 2. | [ROOT, *they*] | *like bagels with lox* | ARC-LEFT | (*they* $\leftarrow$ *like*) |
| 3. | [ROOT] | *like bagels with lox* | SHIFT | |
| 4. | [ROOT, *like*] | *bagels with lox* | SHIFT | |
| 5. | [ROOT, *like, bagels*] | *with lox* | SHIFT | |
| 6. | [ROOT, *like, bagels, with*] | *lox* | ARC-LEFT | (*with* $\leftarrow$ *lox*) |
| 7. | [ROOT, *like, bagels*] | *lox* | ARC-RIGHT | (*bagels* $\rightarrow$ *lox*) |
| 8. | [ROOT, *like*] | *bagels* | ARC-RIGHT | (*like* $\rightarrow$ *bagels*) |
| 9. | [ROOT] | *like* | ARC-RIGHT | (ROOT $\rightarrow$ *like*) |
| 10. | [ROOT] | $\varnothing$ | DONE | |

Table 10.2: Arc-standard derivation of the unlabeled dependency parse for the input *they like bagels with lox*.

- ARC-RIGHT: creates a new right-facing arc between the item on the top of the stack and the first item in the input buffer; this item is then "popped" to the front of the input buffer, and the arc is added to $A$.

$$(\sigma|i, j|\beta, A) \Rightarrow (\sigma, i|\beta, A \oplus (i, j, r)), \quad [10.22]$$

  where again $r$ is the label of the dependency arc.

Each action has preconditions. The SHIFT action can be performed only when the buffer has at least one element. The ARC-LEFT action cannot be performed when the root node ROOT is on top of the stack, since this node must be the root of the entire tree. The ARC-LEFT and ARC-RIGHT remove the modifier words from the stack (in the case of ARC-LEFT) and from the buffer (in the case of ARC-RIGHT), so it is impossible for any word to have more than one parent. Furthermore, the end state can only be reached when every word is removed from the buffer and stack, so the set of arcs is guaranteed to constitute a spanning tree. An example arc-standard derivation is shown in Table 10.2.

**Arc-eager dependency parsing**

In the arc-standard transition system, a word is completely removed from the parse once it has been made the modifier in a dependency arc. At this time, any dependents of this word must have already been identified. Right-branching structures are common in English (and many other languages), with words often modified by units such as prepositional phrases to their right. In the arc-standard system, this means that we must first shift all the units of the input onto the stack, and then work backwards, creating a series of arcs, as occurs in Table 10.2. Note that the decision to shift *bagels* onto the stack guarantees

| | $\sigma$ | $\beta$ | action | arc added to $\mathcal{A}$ |
|---|---|---|---|---|
| 1. | [ROOT] | *they like bagels with lox* | SHIFT | |
| 2. | [ROOT, *they*] | *like bagels with lox* | ARC-LEFT | (*they* $\leftarrow$ *like*) |
| 3. | [ROOT] | *like bagels with lox* | ARC-RIGHT | (ROOT $\rightarrow$ *like*) |
| 4. | [ROOT, *like*] | *bagels with lox* | ARC-RIGHT | (*like* $\rightarrow$ *bagels*) |
| 5. | [ROOT, *like*, *bagels*] | *with lox* | SHIFT | |
| 6. | [ROOT, *like*, *bagels*, *with*] | *lox* | ARC-LEFT | (*with* $\leftarrow$ *lox*) |
| 7. | [ROOT, *like*, *bagels*] | *lox* | ARC-RIGHT | (*bagels* $\rightarrow$ *lox*) |
| 8. | [ROOT, *like*, *bagels*, *lox*] | $\varnothing$ | REDUCE | |
| 9. | [ROOT, *like*, *bagels*] | $\varnothing$ | REDUCE | |
| 10. | [ROOT, *like*] | $\varnothing$ | REDUCE | |
| 11. | [ROOT] | $\varnothing$ | DONE | |

Table 10.3: Arc-eager derivation of the unlabeled dependency parse for the input *they like bagels with lox*.

that the prepositional phrase *with lox* will attach to the noun phrase, and that this decision must be made before the prepositional phrase is itself parsed. This has been argued to be cognitively implausible (Abney and Johnson, 1991); from a computational perspective, it means that a parser may need to look several steps ahead to make the correct decision.

**Arc-eager dependency parsing** changes the ARC-RIGHT action so that right dependents can be attached before all of their dependents have been found. Rather than removing the modifier from both the buffer and stack, the ARC-RIGHT action pushes the modifier on to the stack, on top of the head. Because the stack can now contain elements that already have parents in the partial dependency graph, two additional changes are necessary:

- A precondition is required to ensure that the ARC-LEFT action cannot be applied when the top element on the stack already has a parent in $A$.
- A new REDUCE action is introduced, which can remove elements from the stack if they already have a parent in $A$:

$$(\sigma|i, \beta, A) \Rightarrow (\sigma, \beta, A). \quad [10.23]$$

As a result of these changes, it is now possible to create the arc *like* $\rightarrow$ *bagels* before parsing the prepositional phrase *with lox*. Furthermore, this action does not imply a decision about whether the prepositional phrase will attach to the noun or verb. Noun attachment is chosen in Table 10.3, but verb attachment could be achieved by applying the REDUCE action at step 5 or 7.

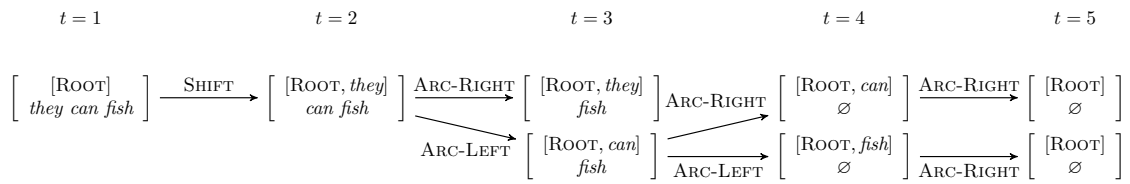



Figure 10.7: Beam search for unlabeled dependency parsing, with beam size $K = 2$. The arc lists for each configuration are not shown, but can be computed from the transitions.

### Projectivity

The arc-standard and arc-eager transition systems are guaranteed to produce projective dependency trees, because all arcs are between the word at the top of the stack and the left-most edge of the buffer (Nivre, 2008). Non-projective transition systems can be constructed by adding actions that create arcs with words that are second or third in the stack (Attardi, 2006), or by adopting an alternative configuration structure, which maintains a list of all words that do not yet have heads (Covington, 2001). In **pseudo-projective dependency parsing**, a projective dependency parse is generated first, and then a set of graph transformation techniques are applied, producing non-projective edges (Nivre and Nilsson, 2005).

### Beam search

In "greedy" transition-based parsing, the parser tries to make the best decision at each configuration. This can lead to search errors, when an early decision locks the parser into a poor derivation. For example, in Table 10.2, if ARC-RIGHT were chosen at step 4, then the parser would later be forced to attach the prepositional phrase *with lox* to the verb *likes*. Note that the *likes* $\rightarrow$ *bagels* arc is indeed part of the correct dependency parse, but the arc-standard transition system requires it to be created later in the derivation.

Beam search addresses this issue by maintaining a set of hypothetical derivations, called a beam. At step $t$ of the derivation, there is a set of $k$ hypotheses, each of which is a tuple of a score and a sequence of actions,

$$h_t^{(k)} = \langle \psi_t^{(k)}, A_t^{(k)} \rangle \quad [10.24]$$

Each hypothesis is then "expanded" by considering the set of all valid actions $\mathcal{A}(c_t^{(k)})$. This yields a large set of new hypotheses. For each new hypothesis $A_t^{(k)} \oplus a$, we compute an updated score. The top $k$ hypotheses by this scoring metric are kept, and parsing proceeds to the next step (Zhang and Clark, 2008). Note that beam search requires a scoring function which applies to action *sequences*, rather than individual actions. This issue will be revisited in the next section.

An example of beam search is shown in Figure 10.7, with a beam size of $K = 2$. For the first transition, the only valid action is SHIFT, so there is only one possible configuration at $t = 2$. From this configuration, there are three possible actions. The top two are ARC-RIGHT and ARC-LEFT, and so the resulting hypotheses from these actions are on the beam at $t = 3$. From these configurations, there are three possible actions each, but the best two are expansions of the bottom hypothesis at $t = 3$. Parsing continues until $t = 5$, at which point both hypotheses reach an accepting state. The best-scoring hypothesis is then selected as the parse.

### 10.3.2 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,

$$\hat{a} = \underset{a \in \mathcal{A}(c)}{\text{argmax}}\, \psi(a, c, \boldsymbol{w}; \boldsymbol{\theta}), \tag{10.25}$$

where $\mathcal{A}(c)$ is the set of admissible actions in the current configuration $c$, $\boldsymbol{w}$ is the input, and $\psi$ is a scoring function with parameters $\boldsymbol{\theta}$ (Yamada and Matsumoto, 2003).

A feature-based score can be computed, $\psi(a, c, \boldsymbol{w}) = \boldsymbol{\theta} \cdot \boldsymbol{f}(a, c, \boldsymbol{w})$, using features that may consider any aspect of the current configuration and input sequence. Typical features for transition-based dependency parsing include: the word and part-of-speech of the top element on the stack; the word and part-of-speech of the first, second, and third elements on the input buffer; pairs and triples of words and parts-of-speech from the top of the stack and the front of the buffer; the distance (in tokens) between the element on the top of the stack and the element in the front of the input buffer; the number of modifiers of each of these elements; and higher-order dependency features as described above in the section on graph-based dependency parsing (see, e.g., Zhang and Nivre, 2011).

Parse actions can also be scored by neural networks. For example, Chen and Manning (2014) build a feedforward networks in which the input layer consists of the concatenation of embeddings of the same words and tags that are used in feature-based approaches:

- the top three words on the stack, and the first three words on the buffer;
- the first and second leftmost and rightmost children (dependents) of the top two words on the stack;
- the leftmost and right most grandchildren of the top two words on the stack;
- embeddings of the part-of-speech tags of these words.

Let us call this base layer $\boldsymbol{x}(c, \boldsymbol{w})$, defined as,

$$\begin{aligned} c =& (\sigma, \beta, A) \\ \boldsymbol{x}(c, \boldsymbol{w}) =& [\boldsymbol{v}_{w_{\sigma_1}}, \boldsymbol{v}_{t_{\sigma_1}} \boldsymbol{v}_{w_{\sigma_2}}, \boldsymbol{v}_{t_{\sigma_2}}, \boldsymbol{v}_{w_{\sigma_3}}, \boldsymbol{v}_{t_{\sigma_3}}, \boldsymbol{v}_{w_{\beta_1}}, \boldsymbol{v}_{t_{\beta_1}}, \boldsymbol{v}_{w_{\beta_2}}, \boldsymbol{v}_{t_{\beta_2}}, \ldots], \end{aligned}$$

where $\boldsymbol{v}_{w_{\sigma_1}}$ is the embedding of the first word on the stack, $\boldsymbol{v}_{t_{\beta_2}}$ 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,

$$\boldsymbol{z} = g(\Theta^{(x \to z)} \boldsymbol{x}(c, \boldsymbol{w})) \quad [10.26]$$

$$\psi(a, c, \boldsymbol{w}; \boldsymbol{\theta}) = \Theta_a^{(z \to y)} \boldsymbol{z}, \quad [10.27]$$

where the vector $\boldsymbol{z}$ plays the same role as the features $\boldsymbol{f}(a, c, \boldsymbol{w})$, but is a learned representation. Chen and Manning (2014) use a cubic elementwise activation function, $g(x) = x^3$, so that the hidden layer models products across all triples of input features. The learning algorithm updates the embeddings as well as the parameters of the feedforward network.

### 10.3.3 Learning to parse

Transition-based dependency parsing suffers from a mismatch between the supervision, which comes in the form of dependency trees, and the classifier's prediction space, which is a set of parsing actions. One solution is to create new training data by converting parse trees into action sequences; another is to derive supervision directly from the parsing performance.

#### Oracle-based training

A transition system can be viewed as a function from action sequences (also called **derivations**) to parse trees. The inverse of this function is a mapping from parse trees to derivations, which is called an **oracle**. For the arc-standard and arc-eager parsing system, an oracle can be computed in linear time in the length of the derivation (Kübler et al., 2009, page 32). Note that both the arc-standard and arc-eager transition systems suffer from **spurious ambiguity**: there exist dependency parses for which multiple derivations are possible, such as $1 \leftarrow 2 \rightarrow 3$ (Cohen et al., 2012). The oracle must choose between these different derivations. For example, the algorithm described by Kübler et al. (2009) would first create the left arc $(1 \leftarrow 2)$, and then create the right arc, $(1 \leftarrow 2) \rightarrow 3$; another oracle might begin by shifting twice, resulting 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):

$$\text{p}(a \mid c, \boldsymbol{w}) = \frac{\exp \psi(a, c, \boldsymbol{w}; \boldsymbol{\theta})}{\sum_{a' \in \mathcal{A}(c)} \exp \psi(a', c, \boldsymbol{w}; \boldsymbol{\theta})} \quad [10.28]$$

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{argmax}} \sum_{i=1}^{N} \sum_{t=1}^{|A^{(i)}|} \log \text{p}(a_t^{(i)} \mid c_t^{(i)}, \boldsymbol{w}), \quad [10.29]$$

where $|A^{(i)}|$ is the length of the action sequence $A^{(i)}$.

Recall that beam search requires a scoring function for action sequences. Such a score can be obtained by adding the log-likelihoods (or hinge losses) across all actions in the sequence (Chen and Manning, 2014).

### Global objectives

The objective in Equation 10.29 is **locally-normalized**, in the sense that it is the product of probabilities over individual actions. A similar characterization could be made of non-probabilistic algorithms in which hinge-loss objectives are summed over individual actions. In either case, training on individual actions can be sub-optimal with respect to global performance, due to the **label bias problem** (Lafferty et al., 2001; Andor et al., 2016).

As a stylized example, suppose that a given configuration appears 100 times in the training data, with action $a_1$ as the oracle action in 51 cases, and $a_2$ as the oracle action in the other 49 cases. However, in cases where $a_2$ is correct, choosing $a_1$ results in a cascade of subsequent errors, while in cases where $a_1$ is correct, choosing $a_2$ results in only a single error. A classifier that is trained on a local objective function will learn to always choose $a_1$, but choosing $a_2$ would minimize the overall number of errors.

This observation motivates a global objective, such as the globally-normalized conditional likelihood,

$$\text{p}(A^{(i)} \mid \boldsymbol{w}; \boldsymbol{\theta}) = \frac{\exp \sum_{t=1}^{|A^{(i)}|} \psi(a_t^{(i)}, c_t^{(i)}, \boldsymbol{w})}{\sum_{A' \in \mathbb{A}(\boldsymbol{w})} \exp \sum_{t=1}^{|A'|} \psi(a'_t, c'_t, \boldsymbol{w})}, \quad [10.30]$$

where the denominator sums over the set of all possible action sequences, $\mathbb{A}(\boldsymbol{w})$.[5] In the conditional random field model for sequence labeling (§ 6.5.3), it was possible to compute this sum explicitly, using dynamic programming. In transition-based parsing, this is not

[5] Andor 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.

possible. However, the sum can be approximated using beam search,

$$\sum_{A' \in \mathbb{A}(\boldsymbol{w})} \exp \sum_{t=1}^{|A'|} \psi(a'_t, c'_t, \boldsymbol{w}) \approx \sum_{k=1}^{K} \exp \sum_{t=1}^{|A^{(k)}|} \psi(a_t^{(k)}, c_t^{(k)}, \boldsymbol{w}), \quad [10.31]$$

where $A^{(k)}$ is an action sequence on a beam of size $K$. This gives rise to the following loss function,

$$L(\boldsymbol{\theta}) = -\sum_{t=1}^{|A^{(i)}|} \psi(a_t^{(i)}, c_t^{(i)}, \boldsymbol{w}) + \log \sum_{k=1}^{K} \exp \sum_{t=1}^{|A^{(k)}|} \psi(a_t^{(k)}, c_t^{(k)}, \boldsymbol{w}). \quad [10.32]$$

The derivatives of this loss involve expectations with respect to a probability distribution over action sequences on the beam.

***Early update and the incremental perceptron**

When learning in the context of beam search, the goal is to learn a decision function so that the gold dependency parse is always reachable from at least one of the partial derivations on the beam. (The combination of a transition system (such as beam search) and a scoring function for actions is known as a **policy**.) To achieve this, we can make an **early update** as soon as the oracle action sequence "falls off" the beam, even before a complete analysis is available (Collins and Roark, 2004; Daumé III and Marcu, 2005). The loss can be based on the best-scoring hypothesis on the beam, or the sum of all hypotheses (Huang et al., 2012).

For example, consider the beam search in Figure 10.7. In the correct parse, *fish* is the head of dependency arcs to both of the other two words. In the arc-standard system, this can be achieved only by using SHIFT for the first two actions. At $t = 3$, the oracle action sequences has fallen off the beam. The parse should therefore stop, and update the parameters by the gradient $\frac{\partial}{\partial \boldsymbol{\theta}} L(A_{1:3}^{(i)}, \{A_{1:3}^{(k)}\}; \boldsymbol{\theta})$, where $A_{1:3}^{(i)}$ is the first three actions of the oracle sequence, and $\{A_{1:3}^{(k)}\}$ is the beam.

This integration of incremental search and learning was first developed in the **incremental perceptron** (Collins and Roark, 2004). This method updates the parameters with respect to a hinge loss, which compares the top-scoring hypothesis and the gold action sequence, up to the current point $t$. This approach can be applied to any discriminatively-trained model. Several improvements to this basic protocol are possible:

- As noted earlier, the gold dependency parse can be derived by multiple action sequences. Rather than checking for the presence of a single oracle action sequence on the beam, we can check if the gold dependency parse is *reachable* from the current beam, using a **dynamic oracle** (Goldberg and Nivre, 2012).

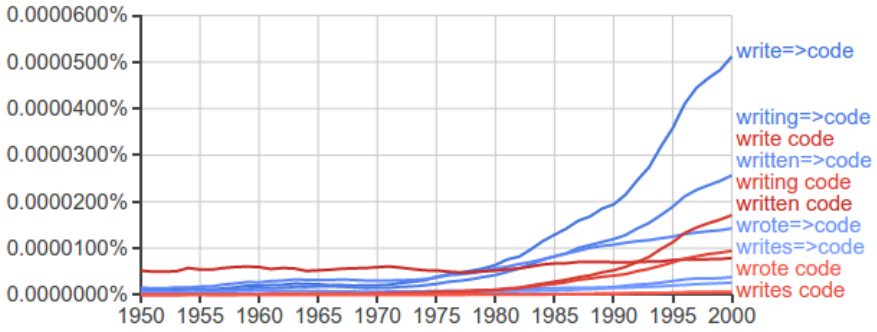



Figure 10.8: Google n-grams results for the bigram *write code* and the dependency arc *write* => *code* (and their morphological variants)

- By maximizing the score of the gold action sequence, we are training a decision function to find the correct action given the gold context. But in reality, the parser will make errors, and the parser is not trained to find the best action given a context that may not itself be optimal. This issue is addressed by various generalizations of incremental perceptron, known as **learning to search** (Daumé III et al., 2009). Some of these methods are discussed in chapter 14.

## 10.4 Applications

Dependency parsing is used in many real-world applications: any time you want to know about pairs of words which might not be adjacent, you can use dependency arcs instead of regular expression search patterns. For example, we may want to match strings like *delicious pastries*, *delicious French pastries*, and *the pastries are delicious*[6]

It is possible to search the Google n-grams corpus by dependency edges, finding the trend in how often a dependency edge appears over time. For example, we might be interested in knowing when people started talking about *writing code*, but we also want *write some code*, *write the code*, *write all the code*, etc. The result of a search on the dependency edge *write* $\rightarrow$ *code* is shown in Figure 10.8. This capability has been applied to research in digital humanities, such as the analysis of Shakespeare performed by Muralidharan and Hearst (2013).

A classic application of dependency parsing is **relation extraction**, which is described

[6]Recall that the copula (in this case, *are*) is collapsed in many dependency grammars, such as the Universal Dependency treebank Nivre et al. (2016).

in chapter 16. The goal of relation extraction is to identify entity pairs, such as

⟨TOLSTOY, WAR AND PEACE⟩
⟨MARQUÉZ, 100 YEARS OF SOLITUDE⟩
⟨SHAKESPEARE, A MIDSUMMER NIGHT'S DREAM⟩,

which stand in some relation to each other (in this case, the relation is authorship). Such entity pairs are often referenced via consistent chains of dependency relations. Therefore, dependency paths are often a useful feature in supervised systems which learn to detect new instances of a relation, based on labeled examples of other instances of the same relation type (Culotta and Sorensen, 2004; Fundel et al., 2007; Mintz et al., 2009).

Cui et al. (2005) show how dependency parsing can improve automated question answering. Suppose you receive the following query:

(10.1) What percentage of the nation's cheese does Wisconsin produce?

The corpus contains this sentence:

(10.2) In Wisconsin, where farmers produce 28% of the nation's cheese, ...

The location of *Wisconsin* in the surface form of this string makes it a poor match for the query. However, in the dependency graph, there is an edge from *produce* to *Wisconsin* in both the question and the potential answer, raising the likelihood that this span of text is relevant to the question.

A final example comes from sentiment analysis. As discussed in chapter 3, the polarity of a sentence can be reversed by negation, e.g.

(10.3) *There is no reason at all to believe the polluters will suddenly become reasonable.*

By tracking the sentiment polarity through the dependency parse, we can better identify the overall polarity of the sentence, determining when key sentiment words are reversed (Wilson et al., 2005; Nakagawa et al., 2010).

## 10.5 Additional reading

More details on dependency grammar and parsing algorithms can be found in the manuscript by Kübler et al. (2009). Jurafsky and Martin (2018) describe an **agenda-based** version of beam search, in which the beam contains hypotheses of varying lengths. New hypotheses are added to the beam only if their score is better than the worst item currently on the beam. Another search algorithm for transition-based parsing is **easy-first**, which abandons the left-to-right traversal order, and adds the highest-scoring edges first, regardless

of where they appear (Goldberg and Elhadad, 2010). Goldberg et al. (2013) note that although transition-based methods can be implemented in linear time in the length of the input, naïve implementations of beam search will require quadratic time, due to the cost of copying each hypothesis when it is expanded on the beam. This issue can be addressed, using a more efficient data structure for the stack.

## Exercises

1. Provide the UD-style dependency parse for the sentence *Xi-Lan eats shoots and leaves*, assuming *leaves* is a verb. Provide arc-standard and arc-eager derivations for this dependency parse.

# Part III

# Meaning

# Chapter 11

# Logical semantics

A grand ambition of natural language processing is to convert natural language into a representation that supports inferences about meaning. Indeed, many potential applications of language technology involve some level of semantic understanding:

- Answering questions, such as *where is the nearest coffeeshop?* or *what is the middle name of the mother of the 44th President of the United States?*.
- Building a robot that can follow natural language instructions to execute tasks.
- Translating a sentence from one language into another, while preserving the underlying meaning.
- Fact-checking an article by searching the web for contradictory evidence.
- Logic-checking an argument by identifying contradictions, ambiguity, and unsupported assertions.

Each of these applications can be performed by converting natural language into a **meaning representation**. To be useful, a meaning representation must meet several criteria:

- **c1**: it should be unambiguous (unlike natural language);
- **c2**: it should provide a way to link language to external knowledge, observations, and actions;
- **c3**: it should support computational **inference**.

Much more than this can be said about the question of how best to represent knowledge for computation (e.g., Sowa, 2000), but we will focus on these three criteria.

## 11.1 Meaning and denotation

The first criterion for a meaning representation is that statements in the representation should be unambiguous — they should have only one possible interpretation. Natural language does not have this property: as we saw in chapter 9, sentences like *cats scratch people with claws* have multiple interpretations.

But what does it mean for a statement to be unambiguous? Programming languages provide a useful example: the output of a program is completely specified by the rules of the language, and the properties of the environment in which the program is run. For example, the python code `5 + 3` will have the output 8, as will the codes `(4*4)-(3*3)+1` and `((8))`. This output is known as the **denotation** of the program, and can be written as,

$$[\![\texttt{5+3}]\!] = [\![\texttt{(4*4)-(3*3)+1}]\!] = [\![\texttt{((8))}]\!] = 8. \quad [11.1]$$

In this sense, the program's output is its "meaning".

The denotations of these arithmetic expressions are determined by the meaning of the **constants** (e.g., `5`, `3`) and the **relations** (e.g., `+`, `*`, `(`, `)`). Now let's consider another snippet of python code, `double(4)`. The denotation of this code could be, $[\![\texttt{double(4)}]\!] = 8$, or it could be $[\![\texttt{double(4)}]\!] = 44$ — it depends on the meaning of `double`. This meaning is defined in a **world model** $\mathcal{M}$ as an infinite set of pairs. We write the denotation with respect to model $\mathcal{M}$ as $[\![\cdot]\!]_{\mathcal{M}}$, e.g., $[\![\texttt{double}]\!]_{\mathcal{M}} = \{\langle \texttt{0}, \texttt{0}\rangle, \langle \texttt{1}, \texttt{2}\rangle, \langle \texttt{2}, \texttt{4}\rangle, \ldots\}$. The world model would also define the (infinite) list of constants, e.g., $\{\texttt{0}, \texttt{1}, \texttt{2}, \ldots\}$. As long as the denotation of string $\phi$ in model $\mathcal{M}$ can be computed unambiguously, the language can be said to be unambiguous.

This approach to meaning is known as **model-theoretic semantics**, and it addresses not only criterion $c1$ (no ambiguity), but also $c2$ (connecting language to external knowledge, observations, and actions). For example, we can connect a representation of the meaning of a statement like *the capital of Georgia* with a world model that includes knowledge base of geographical facts, obtaining the denotation `Atlanta`. We might populate a world model by applying an image analysis algorithm to Figure 11.1, and then use this world model to evaluate **propositions** like *a man is riding a moose*. Another desirable property of model-theoretic semantics is that when the facts change, the denotations change too: the meaning representation of *President of the USA* would have a different denotation in the model $\mathcal{M}_{2014}$ as it would in $\mathcal{M}_{2022}$.

## 11.2 Logical representations of meaning

If we can find a meaning representation which supports model-theoretic denotation, then we will have met criteria $c1$ and $c2$. The final criterion is $c3$, which requires that the meaning representation support inference — for example, automatically deducing new

Figure 11.1: A (doctored) image, which could be the basis for a world model [todo: the image is from 1912, so out of copyright? https://blogs.harvard.edu/houghton/2013/09/20/myths-debunked-sadly-theodore-roosevelt-never-rode-a-moose/]

facts from known premises. While many representations have been proposed that meet these criteria, the most mature is the language of first-order logic.[1]

### 11.2.1 Propositional logic

The bare bones of logical meaning representation are Boolean operations on propositions:

**Propositional symbols** Greek symbols like $\phi$ and $\psi$ will be used to represent **propositions**, which are statements that are either true or false. For example, $\phi$ may correspond to the proposition, *bagels are delicious*.

**Boolean operators** We can build up more complex propositional formulas from Boolean operators. These include:

- Negation $\neg\phi$, which is true if $\phi$ is false.
- Conjunction, $\phi \wedge \psi$, which is true if both $\phi$ and $\psi$ are true.

[1] Relevant alternatives 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).

- Disjunction, $\phi \vee \psi$, which is true if at least one of $P$ and $Q$ is true
- Implication, $\phi \Rightarrow \psi$, which is true unless $\phi$ is true and $\psi$ is false. Implication has identical truth conditions to $\neg\phi \vee \psi$.
- Equivalence, $\phi \Leftrightarrow \psi$, which is true if $\phi$ and $\psi$ are both true or both false. Equivalence has identical truth conditions to $(\phi \Rightarrow \psi) \wedge (\psi \Rightarrow \phi)$.

It is not strictly necessary to have all five Boolean operators: readers familiar with Boolean logic will know that it is possible to construct all other operators from either the NAND (not-and) or NOR (not-or) operators. Nonetheless, it is typical to use all five operators above for clarity. It is possible to define a number of "laws" for these Boolean operators, such as,

- **commutativity**: $\phi \wedge \psi = \psi \wedge \phi, \quad \phi \vee \psi = \psi \vee \phi$
- **associativity**: $\phi \wedge (\psi \wedge \chi) = (\phi \wedge \psi) \wedge \chi, \quad \phi \vee (\psi \vee \chi) = (\phi \vee \psi) \vee \chi$
- **complementation**: $\phi \wedge \neg\phi = \bot, \quad \phi \vee \neg\phi = \top$, where $\top$ indicates a true proposition and $\bot$ indicates a false proposition.

These laws can be combined to derive further equivalences, which can support logical inferences. For example, suppose $\phi =$ *The music is loud* and $\psi =$ *Max can't sleep*. Then if we are given,

$$\begin{aligned} \phi \Rightarrow \psi \quad & \textit{If the music is loud, Max can't sleep.} \\ \phi \quad & \textit{The music is loud.} \end{aligned}$$

we can derive $\psi$ (*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).

### 11.2.2 First-order logic

Propositional logic is so named because it treats propositions as its base units. However, we would also like to reason about the content of the propositions themselves. For example,

(11.1) If anyone is making noise, then Max can't sleep.

(11.2) Abigail is making noise.

To understand the relationship between the statement *anyone is making noise* and the statement *Abigail is making noise*, we need some additional formal machinery. This is provided by **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., $[\![\text{MAX}]\!] = \texttt{m}$ and $[\![\text{ABIGAIL}]\!] = \texttt{a}$.

**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 $[\![\text{CAN-SLEEP}]\!] = \{\texttt{a}, \texttt{m}, \ldots\}$. We can test the truth value of the proposition CAN-SLEEP(MAX) by asking 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, $[\![\text{BROTHER}]\!] = \{(\texttt{m}, \texttt{a}), (\texttt{x}, \texttt{y}), \ldots\}$. We can test the truth value of the proposition BROTHER(MAX,ABIGAIL) by asking whether the tuple $([\![\text{MAX}]\!], [\![\text{ABIGAIL}]\!])$ is in the denotation $[\![\text{BROTHER}]\!]$.

We can now express statements like *Max can't sleep* and *Max is Abigail's brother*:

$$\neg\text{CAN-SLEEP}(\text{MAX})$$
$$\text{BROTHER}(\text{MAX},\text{ABIGAIL}).$$

We can also combine these statements using Boolean operators, such as,

$$(\text{BROTHER}(\text{MAX},\text{ABIGAIL}) \vee \text{BROTHER}(\text{MAX},\text{STEVE})) \Rightarrow \neg\text{CAN-SLEEP}(\text{MAX}).$$

We have thus far described a fragment of first-order logic, which permits only statements about specific entities. To support inferences about statements like *If **anyone** is making noise, then Max can't sleep*, we will need two additional elements in the meaning representation:

**Variables** These are mechanisms for referring to entities that are not locally specified. We can then write CAN-SLEEP$(x)$ or BROTHER$(x, \text{ABIGAIL})$. In these cases, $x$ is an **free variable**, meaning that we have not committed to any particular assignment.

**Quantifiers** Variables are bound by quantifiers. There are two quantifiers in first-order logic.[2]

[2]In first-order logic, it is possible to quantify only over entities. In **second-order logic**, it is possible to quantify over properties, supporting statements like *Butch has every property that a good boxer has*,

$$\forall P \forall x((\text{GOOD-BOXER}(x) \Rightarrow P(x)) \Rightarrow P(\text{BUTCH}) \quad [11.2]$$

This example is from Blackburn and Bos (2005), who also show how first-order logic can "approximate" second-order and higher-order logics, by reifying sets as additional entities in the model.

- The **existential quantifier** $\exists$, which indicates that there must be at least one entity to which the variable can refer. For example, the statement $\exists x$MAKES-NOISE(X) indicates that there is at least one entity for which MAKES-NOISE is true.
- The **universal quantifier** $\forall$, which indicates that the the variable must be able to refer to any entity. For example, the statement,

$$\neg\text{MAKES-NOISE}(\text{ABIGAIL}) \Rightarrow (\forall x \text{CAN-SLEEP}(x)) \quad [11.3]$$

  asserts that every entity can sleep if Abigail does not make noise.

The expressions $\exists x$ and $\forall x$ make $x$ into a **bound variable**. A formula that contains no free variables is a **sentence**.

**Functions** Functions map from entities to entities, e.g., $[\![\text{CAPITAL-OF}(\text{GEORGIA})]\!] = [\![\text{ATLANTA}]\!]$. With functions, we also add an equality operator, so that it is possible to make statements like,

$$\forall x \exists y \text{MOTHER-OF}(x) = \text{DAUGHTER-OF}(y). \quad [11.4]$$

Note that MOTHER-OF is a functional analogue of the relation MOTHER, so that MOTHER-OF$(x) = y$ if MOTHER$(x, y)$. Any logical formula that uses functions can be rewritten using only relations and quantification. For example,

$$\text{MAKES-NOISE}(\text{MOTHER-OF}(\text{ABIGAIL})) \quad [11.5]$$

can be rewritten as $\exists x$MAKES-NOISE$(x) \wedge$ MOTHER$(x,$ 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:

$$\forall x \exists y \text{ SPEAKS}(x, y) \quad [11.6]$$
$$\exists y \forall x \text{ SPEAKS}(x, y). \quad [11.7]$$

In the first case, $y$ may refer to several different languages, while in the second case, there is a single $y$ that is spoken by everyone.

### Truth-conditional semantics

One way to look at the meaning of an FOL sentence $\phi$ is as a set of **truth conditions**, or models under which $\phi$ is satisfied. But how can we determine whether a sentence in first-order logic is true or false? We will approach this inductively, starting with a predicate applied to a tuple of constants. The truth of such a sentence depends on whether the

tuple of denotations of the constants is in the denotation of the predicate. For example, CAPITAL(GEORGIA,ATLANTA) is true in model $\mathcal{M}$ iff,

$$\langle [\![\text{GEORGIA}]\!]_{\mathcal{M}}, [\![\text{ATLANTA}]\!]_{\mathcal{M}} \rangle \in [\![\text{CAPITAL}]\!]_{\mathcal{M}}. \quad [11.8]$$

The Boolean operators $\wedge, \vee, \ldots$ provide ways to construct more complicated sentences, and the truth of such statements can be assessed based on the truth tables associated with these operators. The statement $\exists x \phi$ is true if there is some assignment of the variable $x$ to an entity in the model such that $\phi$ is true; the statement $\forall x \phi$ is true if $\phi$ is true under all possible assignments of $x$. More formally, we would say that $\phi$ is **satisfied** under $\mathcal{M}$, written as $\mathcal{M} \models \phi$.

Truth conditional semantics allows us to define several other properties of sentences and pairs of sentences. Suppose that in every $\mathcal{M}$ under which $\phi$ is satisfied, another formula $\psi$ is also satisfied; then $\phi$ **entails** $\psi$, sometimes written $\phi \models \psi$ [todo: double check]. For example,

$$\text{CAPITAL(GEORGIA,ATLANTA)} \models \exists x \text{CAPITAL}(\text{GEORGIA}, x). \quad [11.9]$$

A statement that is satisfied under any model, such as $\phi \vee \neg\phi$, is **valid**; a statement that is not satisfied under any model, such as $\phi \wedge \neg\phi$, is **unsatisfiable**, or **inconsistent**. A **model checker** is a program that determines whether a sentence $\phi$ is satisfied in $\mathcal{M}$. A **model builder** is a program that constructs a model in which $\phi$ is satisfied. The problems of checking for consistency and validity in first-order logic are **undecidable**, meaning that there is no algorithm that can automatically determine whether an FOL formula is valid or inconsistent.

#### Inference in first-order logic

Our original goal was to support inferences that combine general statements *If anyone is making noise, then Max can't sleep* with specific statements like *Abigail is making noise*. We can now represent such statements in first-order logic, but how are we to perform the inference that *Max can't sleep*? One approach is to use "generalized" versions of propositional inference rules like modus ponens, which can be applied to FOL formulas. By repeatedly applying such inference rules to a knowledge base of facts, it is possible to produce proofs of desired propositions. To find the right sequence of inferences to derive a desired theorem, classical artificial intelligence search algorithms like backward chaining can be applied. Such algorithms are implemented in interpreters for the `prolog` logic programming language (Pereira and Shieber, 2002).

## 11.3 Semantic parsing and the lambda calculus

The previous section has laid out a lot of formal machinery, which we will now try to unite with natural language. Given an English sentence like *Alex likes Brit*, how can we

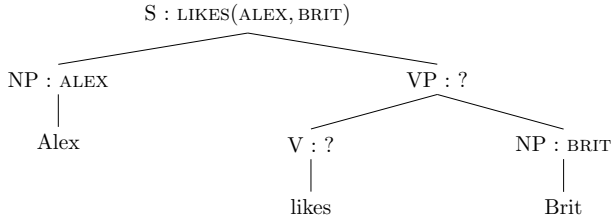



Figure 11.2: The principle of compositionality requires that we identify meanings for the constituents *likes* and *likes Brit* that will make it possible to compute the meaning for the entire sentence.

obtain the desired first-order logical representation, LIKES(ALEX,BRIT)? This is the task of **semantic parsing**. Just as a syntactic parser is a function from a natural language sentence to a syntactic structure such as a phrase structure tree, a semantic parser is a function from natural language to logical formulas.

As in syntactic analysis, semantic parsing is difficult because the space of inputs and outputs is very large, and their interaction is complex. Our best hope is that, like syntactic parsing, semantic parsing can somehow be decomposed into simpler sub-problems. This idea, usually attributed to the German philosopher Gottlob Frege, is called the **principle of compositionality**: the meaning of a complex expression is a function of the meanings of that expression's constituent parts. We will define these "constituent parts" as syntactic elements like noun phrases and verb phrases. These constituents are combined using function application: if the syntactic parse contains the production $x \rightarrow y\ z$, then the semantics of $x$, written $x$.sem, will be computed as a function of the semantics of the constituents, $y$.sem and $z$.sem.[3] [4]

### 11.3.1 The lambda calculus

Let's see how this works for a simple sentence like *Alex likes Brit*, whose syntactic structure is shown in Figure 11.2. Our goal is the formula, LIKES(ALEX,BRIT), and it is clear that the meaning of the constituents *Alex* and *Brit* should be ALEX and BRIT. That leaves two more constituents: the verb *likes*, and the verb phrase *likes Brit*. How can we define the meaning of these units in a way that enables us to recover the desired meaning for the

[3] § 8.3.1 briefly discusses alternative syntactic formalisms, including Combinatory Categorial Grammar (CCG). 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 § 11.4.

[4] The 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).

entire sentence? If the meanings of *Alex* and *Brit* are constants, then the meanings of *likes* and *likes Brit* must be functional expressions, which can be applied to their siblings to produce the desired analyses.

Modeling these partial analyses requires extending the first-order logic meaning representation. We do this by adding **lambda expressions**, which are descriptions of anonymous functions,[5] e.g.,

$$\lambda x.\text{LIKES}(x, \text{BRIT}). \quad [11.10]$$

We take this functional expression to be the meaning of the verb phrase *likes Brit*; it takes a single argument, and returns the result of substituting that argument for $x$ in the expression $\text{LIKES}(x, \text{BRIT})$. We write this substitution as,

$$(\lambda x.\text{LIKES}(x, \text{BRIT}))@\text{ALEX} = \text{LIKES}(\text{ALEX},\text{BRIT}), \quad [11.11]$$

with the symbol "@" indicating function application. Function application in the lambda calculus is sometimes called $\beta$**-reduction** or $\beta$**-conversion**. We will write $\phi@\psi$ to indicate a function application to be performed by $\beta$-reduction, and $\phi(\psi)$ to indicate a function or predicate in the final logical form.

Equation 11.11 shows how to obtain the desired semantics for the sentence *Alex likes Brit*: by applying the lambda expression $\lambda x.\text{LIKES}(x, \text{BRIT})$ to the logical constant ALEX. This rule of composition can be specified in a syntactic-semantic grammar: for the syntactic production S $\rightarrow$ NP VP, we have the semantic rule VP.sem@NP.sem.

The meaning of the meaning of the transitive verb phrase can also be obtained by function application on its syntactic constituents. For the syntactic production VP $\rightarrow$ V NP, we apply the semantic rule,

$$\begin{aligned} \text{VP.sem} &= (\text{V.sem})@\text{NP.sem} && [11.12] \\ &= (\lambda y.\lambda x.\text{LIKES}(x, y))@(\text{BRIT}) && [11.13] \\ &= \lambda x.\text{LIKES}(x, \text{BRIT}). && [11.14] \end{aligned}$$

Here we have defined the meaning of the transitive verb *likes* as a lambda expression whose output is **another** lambda expression: it takes $y$ as an argument to fill in one of the slots in the LIKES relation, and returns a lambda expression that is ready to take an argument to fill in the other slot.[6]

[5] Formally, all first-order logic formulas are lambda expressions; in addition, if $\phi$ 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.

[6] This can be written in a few different ways. More informally, we can write $\lambda y, x.\text{LIKES}(x, y)$, indicating a lambda expression that takes two arguments; this would be acceptable in functional programming. More formally, logicians (e.g., Carpenter, 1997) often write $\lambda y.\lambda x.\text{LIKES}(x)(y)$, indicating that each lambda expression takes exactly one argument.

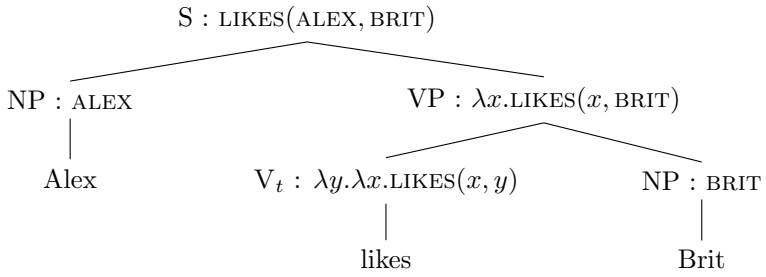



Figure 11.3: Derivation of the semantic representation for *Alex likes Brit* in the grammar $G_1$.

| | | |
|---|---|---|
| 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 11.1: $G_1$, a minimal syntactic/semantic context-free grammar

Table 11.1 shows a minimal syntactic/semantic grammar fragment, which we will call $G_1$. The complete **derivation** of *Alex likes Brit* in $G_1$ is shown in Figure 11.3. In addition to the transitive verb *likes*, the grammar also includes the intransitive verb *sleeps*; it should be clear how to derive the meaning of sentences like *Alex sleeps*. For verbs that can be either transitive or intransitive, such as *eats*, we would have two terminal productions, one for each sense (terminal productions are also called the **lexical entries**). Indeed, most of the grammar is in the **lexicon** (the terminal productions), since these productions select the basic units of the semantic interpretation.

### 11.3.2 Quantification

Things get more complicated when we move from sentences about named entities to sentences that involve more general noun phrases. Let's consider the example, *A dog sleeps*, 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]

[7]Conversely, 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

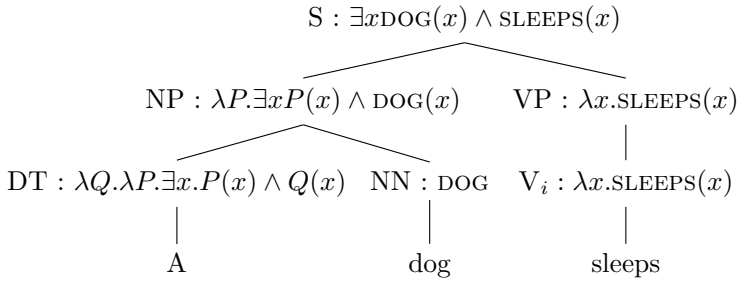



Figure 11.4: Derivation of the semantic representation for *A dog sleeps*, in grammar $G_2$

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 $\rightarrow$ NP VP has the semantics NP.sem@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 11.4. Let's start with the indefinite article *a*, to which we assign the rather intimidating semantics,

$$\lambda P.\lambda Q.\exists x P(x) \wedge Q(x). \quad [11.15]$$

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,

$$\begin{aligned} \text{NP.sem} &= \text{DET.sem@NN.sem} && [11.16] \\ &= (\lambda P.\lambda Q.\exists x P(x) \wedge Q(x))@(\text{DOG}) && [11.17] \\ &= \lambda Q.\exists x \text{DOG}(x) \wedge Q(x). && [11.18] \end{aligned}$$

This is a lambda expression that is expecting another relation, $Q$, which will be provided by the verb phrase, SLEEPS. This gives the desired analysis, $\exists x \text{DOG}(x) \wedge \text{SLEEPS}(x)$.[9]

identifiable, perhaps from contextual information external to the sentence. Carpenter (1997, pp. 96-100) summarizes recent approaches to handling definite descriptions.

[8] Carpenter (1997) offers an alternative treatment based on combinatory categorial grammar.

[9] When applying $\beta$-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 $\alpha$**-conversion**. For example, $\lambda x.P(x)$ can be converted to $\lambda y.P(y)$, etc.

If noun phrases like *a dog* are interpreted as lambda expressions, then proper nouns like *Alex* must be treated in the same way. This is achieved by **type-raising** from constants to lambda expressions, $x \Rightarrow \lambda P.P(x)$. After type-raising, the semantics of *Alex* is $\lambda P.P(\text{ALEX})$ — a lambda expression that expects a relation to tell us something about ALEX.[10] Make sure you see how the analysis in Figure 11.4 can be applied to the sentence *Alex sleeps*.

To handle direct objects, we will perform the same type-raising operation on transitive verbs: the meaning of verbs such as *likes* will be raised to,

$$\lambda P.\lambda x.P(\lambda y.\text{LIKES}(x, y)) \quad [11.19]$$

As a result, we can keep the verb phrase production VP.sem = V.sem@NP.sem, knowing that the direct object will provide the function $P$ in Equation 11.19. To see how this works, let's analyze the verb phrase *likes a dog*. After uniquely relabeling each lambda variable, we have,

$$\begin{aligned}
\text{VP.sem} =& \text{V.sem@NP.sem} \\
=& (\lambda P.\lambda x.P(\lambda y.\text{LIKES}(x, y)))@(\lambda Q.\exists z \text{DOG}(z) \wedge Q(z)) \\
=& \lambda x.(\lambda Q.\exists z \text{DOG}(z) \wedge Q(z))@(\lambda y.\text{LIKES}(x, y)) \\
=& \lambda x.\exists z \text{DOG}(z) \wedge (\lambda y.\text{LIKES}(x, y))@z \\
=& \lambda x.\exists z \text{DOG}(z) \wedge \text{LIKES}(x, z).
\end{aligned}$$

These changes are summarized in the revised grammar $G_2$, shown in Table 11.2. Figure 11.5 shows a derivation that involves a transitive verb, an indefinite noun phrase, and a proper noun.

## 11.4 Learning semantic parsers

As with syntactic parsing, any syntactic/semantic grammar with sufficient coverage will **overgenerate**, producing many possible analyses for any given sentence. Machine learning is the dominant approach to selecting a single analysis. We will focus on algorithms that learn to score logical forms by attaching weights to features of their derivations (Zettlemoyer and Collins, 2005). Alternative approaches include transition-based parsing (Zelle

[10] Compositional 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.

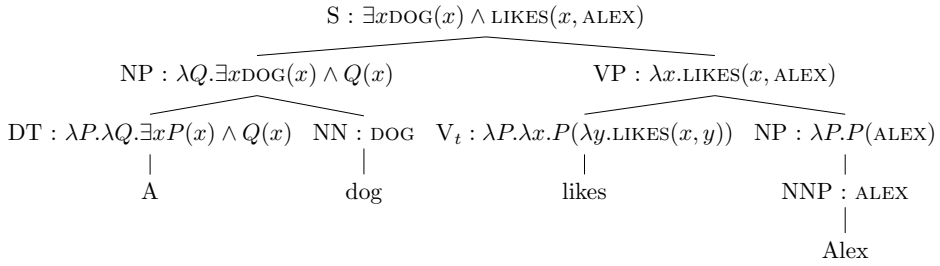



Figure 11.5: Derivation of the semantic representation for *A dog likes Alex*.

| | | |
|---|---|---|
| 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* | $\lambda P.\lambda Q.\exists x P(x) \wedge Q(x)$ |
| DET | $\rightarrow$ *every* | $\lambda P.\lambda Q.\forall x(P(x) \Rightarrow Q(x))$ |
| $V_t$ | $\rightarrow$ *likes* | $\lambda P.\lambda x.P(\lambda y.\text{LIKES}(x, y))$ |
| $V_i$ | $\rightarrow$ *sleeps* | $\lambda x.\text{SLEEPS}(x)$ |
| NN | $\rightarrow$ *dog* | DOG |
| NNP | $\rightarrow$ *Alex* | ALEX |
| NNP | $\rightarrow$ *Brit* | BRIT |

Table 11.2: $G_2$, a syntactic/semantic context-free grammar fragment, which supports quantified noun phrases

and Mooney, 1996; Misra and Artzi, 2016) and methods inspired by machine translation (Wong and Mooney, 2006). Methods also differ in the form of supervision used for learning, which can range from complete derivations to much more limited training signals. We will begin with the case of complete supervision, and then consider how learning is still possible even when seemingly key information is missing.

**Datasets** Early work on semantic parsing focused on geographical queries, such as *What states border Texas*. The GeoQuery dataset of Zelle and Mooney (1996) was originally coded in prolog, but has subsequently been expanded and converted into the SQL database query language by Popescu et al. (2003) and into first-order logic with lambda calculus by Zettlemoyer and Collins (2005), providing logical forms like $\lambda x.\text{STATE}(x) \wedge \text{BORDERS}(x, \text{TEXAS})$. Another early dataset consists of instructions for RoboCup robot soccer teams (Kate et al., 2005). More recent work has focused on broader domains, such as the Freebase database (Bol-

lacker et al., 2008), for which queries have been annotated by Krishnamurthy and Mitchell (2012) and Cai and Yates (2013), as well on child-directed speech (Kwiatkowski et al., 2012) and elementary school science exams (Krishnamurthy, 2016).

### 11.4.1 Learning from derivations

Let $\boldsymbol{w}^{(i)}$ indicate a sequence of text, and let $\boldsymbol{y}^{(i)}$ indicate the desired logical form. For example:

$$\boldsymbol{w}^{(i)} = \text{Alex eats shoots and leaves}$$
$$\boldsymbol{y}^{(i)} = \text{EATS(ALEX,SHOOTS)} \wedge \text{EATS(ALEX,LEAVES)}$$

In the standard supervised learning paradigm that was introduced in § 1.2, we first define a feature function, $\boldsymbol{f}(\boldsymbol{w}, \boldsymbol{y})$, and then learn weights on these features, so that $\boldsymbol{y}^{(i)} = \text{argmax}_{\boldsymbol{y}} \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, \boldsymbol{y})$. The weight vector $\boldsymbol{\theta}$ is learned by comparing the features of the true label $\boldsymbol{f}(\boldsymbol{w}^{(i)}, \boldsymbol{y}^{(i)})$ against either the features of the predicted label $\boldsymbol{f}(\boldsymbol{w}^{(i)}, \hat{\boldsymbol{y}})$ (perceptron, support vector machine) or the expected feature vector $E_{\boldsymbol{y}|\boldsymbol{w}}[\boldsymbol{f}(\boldsymbol{w}^{(i)}, \boldsymbol{y})]$ (logistic regression).

While this basic framework seems similar to discriminative syntactic parsing, there is a crucial difference. In (context-free) syntactic parsing, the annotation $\boldsymbol{y}^{(i)}$ 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) ∧ 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, § 10.3.2.)

Let us introduce an additional variable $\boldsymbol{z}$, representing the **derivation** of the logical form $\boldsymbol{y}$ from the text $\boldsymbol{w}$. We assume that the feature function decomposes across the productions in the derivation, $\boldsymbol{f}(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{y}) = \sum_{t=1}^{T} \boldsymbol{f}(\boldsymbol{w}, z_t, \boldsymbol{y})$, where $z_t$ indicates a single syntactic/semantic production. For example, we might have a feature for the high-level production S → NP VP : NP.sem@VP.sem, as well as for terminal productions like $\text{N}_{\text{NP}}$ → *Alex* : ALEX. Under this decomposition, we can compute scores for each semantically-annotated subtree in the analysis of $\boldsymbol{w}$, and can therefore apply bottom-up parsing algorithms like CKY (§ 9.1) to find the best-scoring semantic analysis.

Figure 11.6 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 $\text{LEAVES}_n$. Figure 11.7 shows a derivation of an incorrect analysis. Assuming one feature per production, the perceptron update is shown in Table 11.3. From this update, the parser would learn to prefer the noun

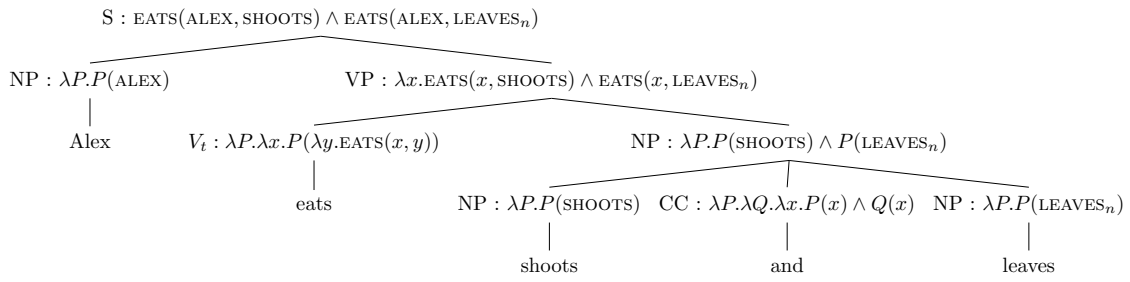



Figure 11.6: Derivation for gold semantic analysis of *Alex eats shoots and leaves*

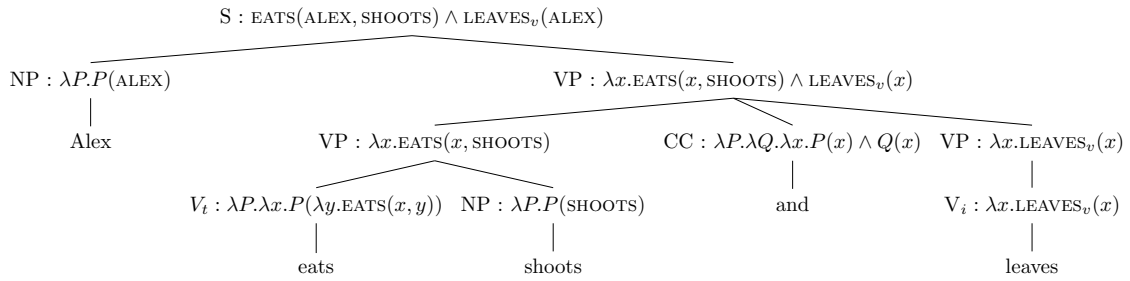



Figure 11.7: Derivation for incorrect semantic analysis of *Alex eats shoots and leaves*

| | | |
|---|---|---|
| $\text{NP}_1 \rightarrow \text{NP}_2$ CC $\text{NP}_3$ | (CC.sem@($\text{NP}_2$.sem))@($\text{NP}_3$.sem) | +1 |
| $\text{VP}_1 \rightarrow \text{VP}_2$ CC $\text{VP}_3$ | (CC.sem@($\text{VP}_2$.sem))@($\text{VP}_3$.sem) | -1 |
| NP $\rightarrow$ *leaves* | LEAVES$_n$ | +1 |
| VP $\rightarrow$ $\text{V}_i$ | $\text{V}_i$.sem | -1 |
| $\text{V}_i \rightarrow$ *leaves* | $\lambda x$.LEAVES$_v$ | -1 |

Table 11.3: Perceptron update for analysis in Figure 11.6 (gold) and Figure 11.7 (predicted)

interpretation of *leaves* over the verb interpretation. It would also learn to prefer noun phrase coordination over verb phrase coordination. Note that we could easily replace the perceptron with a conditional random field. In this case, the online updates would be based on feature expectations, which can be computed using bottom-up algorithms like inside-outside (§ 9.6).

### 11.4.2 Learning from logical forms

Complete derivations are expensive to annotate, and are rarely available.[11] More recent work has focused 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 $\boldsymbol{y}$ and derivations $\boldsymbol{z}$, we have,

$$\text{p}(\boldsymbol{y}, \boldsymbol{z} \mid \boldsymbol{w}) = \frac{\exp(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{y}))}{\sum_{\boldsymbol{y}', \boldsymbol{z}'} \exp(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, \boldsymbol{z}', \boldsymbol{y}'))}, \tag{11.20}$$

which is the standard log-linear model, applied to the logical form $\boldsymbol{y}$ and the derivation $\boldsymbol{z}$.

Since the derivation $\boldsymbol{z}$ completely determines the logical form $\boldsymbol{y}$, it may seem silly to model the joint probability over $\boldsymbol{y}$ and $\boldsymbol{z}$. However, since $\boldsymbol{z}$ is unknown, it can be marginalized out,

$$\text{p}(\boldsymbol{y} \mid \boldsymbol{w}) = \sum_{\boldsymbol{z}} \text{p}(\boldsymbol{y}, \boldsymbol{z} \mid \boldsymbol{w}). \tag{11.21}$$

We can then have the semantic parser select the logical form with the maximum log marginal probability,

$$\log \sum_{\boldsymbol{z}} \text{p}(\boldsymbol{y}, \boldsymbol{z} \mid \boldsymbol{w}) = \log \sum_{\boldsymbol{z}} \frac{\exp(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{y}))}{\sum \boldsymbol{y}', \boldsymbol{z}' \exp(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, \boldsymbol{z}', \boldsymbol{y}'))} \tag{11.22}$$

$$\propto \log \sum_{\boldsymbol{z}} \exp(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, \boldsymbol{z}', \boldsymbol{y}')) \tag{11.23}$$

$$\geq \max_{\boldsymbol{z}} \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{y}). \tag{11.24}$$

Note that it is impossible to push the $\log$ term inside the sum over $\boldsymbol{z}$, meaning that 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 $\boldsymbol{z}$, which provides a lower bound.

Learning can be performed by maximizing the log marginal likelihood,

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log \text{p}(\boldsymbol{y}^{(i)} \mid \boldsymbol{w}^{(i)}; \boldsymbol{\theta}) \tag{11.25}$$

$$= \sum_{i=1}^{N} \log \sum_{\boldsymbol{z}} \text{p}(\boldsymbol{y}^{(i)}, \boldsymbol{z}^{(i)} \mid \boldsymbol{w}^{(i)}; \boldsymbol{\theta}). \tag{11.26}$$

[11] An exception is the work of Ge and Mooney (2005), who annotate the meaning of each syntactic constituents for several hundred sentences.

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 11.26 is,

$$\frac{\partial \ell_i}{\partial \boldsymbol{\theta}} = \sum_{\boldsymbol{z}} \text{p}(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{w}; \boldsymbol{\theta}) \boldsymbol{f}(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{y}) - \sum_{\boldsymbol{y}', \boldsymbol{z}'} \text{p}(\boldsymbol{y}', \boldsymbol{z}' \mid \boldsymbol{w}; \boldsymbol{\theta}) \boldsymbol{f}(\boldsymbol{w}, \boldsymbol{z}', \boldsymbol{y}') \quad [11.27]$$

$$= E_{\boldsymbol{z}|\boldsymbol{y},\boldsymbol{w}} \boldsymbol{f}(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{y}) - E_{\boldsymbol{y},\boldsymbol{z}|\boldsymbol{w}} \boldsymbol{f}(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{y}) \quad [11.28]$$

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 $\boldsymbol{z}$ that are compatible with the logical form $\boldsymbol{y}$. This can be done in a bottom-up dynamic programming algorithm, by having each cell in the table $t[i, j]$ include both the constituent types (e.g., NP, S) that can derive the span $\boldsymbol{w}_{i:j-1}$, as well as the set of possible logical forms. The inclusion of logical forms means that the resulting table may be much larger than in syntactic parsing, where it is limited by the number of non-terminals in the grammar. This can be controlled by using pruning to eliminate intermediate analyses that are incompatible with the final logical form $\boldsymbol{y}$ (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 11.28 with $\text{argmax}$ and then apply stochastic gradient descent to learn the weights, we obtain the **latent variable perceptron**, a simple and general algorithm for learning with missing data. The algorithm is shown in its most basic form in Algorithm 15, 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).

### 11.4.3 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

**Algorithm 15** Latent variable perceptron

**procedure** LATENTVARIABLEPERCEPTRON($\boldsymbol{w}^{(1:N)}, \boldsymbol{y}^{(1:N)}$)
  $\boldsymbol{\theta} \leftarrow \mathbf{0}$
  **repeat**
    Select an instance $i$
    $\boldsymbol{z}^{(i)} \leftarrow \text{argmax}_{\boldsymbol{z}}\, \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}^{(i)}, \boldsymbol{z}, \boldsymbol{y}^{(i)})$
    $\hat{\boldsymbol{y}}, \hat{\boldsymbol{z}} \leftarrow \text{argmax}_{\boldsymbol{y}', \boldsymbol{z}'}\, \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}^{(i)}, \boldsymbol{z}', \boldsymbol{y}')$
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \boldsymbol{f}(\boldsymbol{w}^{(i)}, \boldsymbol{z}^{(i)}, \boldsymbol{y}^{(i)}) - \boldsymbol{f}(\boldsymbol{w}^{(i)}, \hat{\boldsymbol{z}}, \hat{\boldsymbol{y}})$
  **until** tired
  **return** $\boldsymbol{\theta}$

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 $v_i(\boldsymbol{y}) \in \{0, 1\}$ be a **validation function**, which assigns a binary score indicating whether the denotation $[\![\boldsymbol{y}]\!]$ for the text $\boldsymbol{w}^{(i)}$ is correct. We can then learn by maximizing a conditional-likelihood objective,

$$\ell^{(i)}(\boldsymbol{\theta}) = \log \sum_{\boldsymbol{y}} v_i(\boldsymbol{y}) \times \text{p}(\boldsymbol{y} \mid \boldsymbol{w}; \boldsymbol{\theta}) \quad [11.29]$$

$$= \log \sum_{\boldsymbol{y}} v_i(\boldsymbol{y}) \times \sum_{\boldsymbol{z}} \text{p}(\boldsymbol{y}, \boldsymbol{z} \mid \boldsymbol{w}; \boldsymbol{\theta}), \quad [11.30]$$

which sums over all derivations $\boldsymbol{z}$ of all valid logical forms, $\{\boldsymbol{y} : v_i(\boldsymbol{y}) = 1\}$. 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,

$$\frac{\partial \ell^{(i)}}{\partial \boldsymbol{\theta}} = \sum_{\boldsymbol{y}, \boldsymbol{z}: v_i(\boldsymbol{y})=1} \text{p}(\boldsymbol{y}, \boldsymbol{z} \mid \boldsymbol{w}) \boldsymbol{f}(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{y}) - \sum_{\boldsymbol{y}', \boldsymbol{z}'} \text{p}(\boldsymbol{y}', \boldsymbol{z}' \mid \boldsymbol{w}) \boldsymbol{f}(\boldsymbol{w}, \boldsymbol{z}', \boldsymbol{y}'), \quad [11.31]$$

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).

To learn more about ongoing research on data-driven semantic parsing, readers may consult the survey article by Liang and Potts (2015), tutorial slides and videos by Artzi and Zettlemoyer (2013),[12] and the source code by Yoav Artzi[13] and Percy Liang.[14]

## Exercises

1. Derive the **modus ponens** inference rule, which states that if we know $\phi \Rightarrow \psi$ and $\phi$, then $\psi$ must be true. The derivation can be performed using the definition of the $\Rightarrow$ operator and some of the laws provided in § 11.2.1, plus one additional identity: $\bot \vee \phi = \phi$.

2. Convert the following examples into first-order logic, using the relations CAN-SLEEP, MAKES-NOISE, and BROTHER.

   - If Abigail makes noise, no one can sleep.
   - If Abigail makes noise, someone cannot sleep.

[12] Videos are currently available at `http://yoavartzi.com/tutorial/`

[13] `http://yoavartzi.com/spf`

[14] `https://github.com/percyliang/sempre`

- None of Abigail's brothers can sleep.
- If one of Abigail's brothers makes noise, Abigail cannot sleep.

3. Extend the grammar fragment $G_1$ to include the ditransitive verb *teaches* and the proper noun *Swahili*. Show how to derive the interpretation for the sentence *Alex teaches Brit Swahili*, which should be TEACHES(ALEX,BRIT,SWAHILI). The grammar need not be in Chomsky Normal Form. For the ditransitive verb, use $\text{NP}_1$ and $\text{NP}_2$ to indicate the two direct objects.

4. Derive the semantic interpretation for the sentence *Alex likes every dog*, using grammar fragment $G_2$.

5. Extend the grammar fragment $G_2$ to handle adjectives, so that the meaning of *an angry dog* is $\lambda P.\exists x \text{DOG}(x) \wedge \text{ANGRY}(x) \wedge P(x)$. Specifically, you should supply the lexical entry for the adjective *angry*, and you should specify the syntactic-semantic productions NP $\rightarrow$ DET NOM, NOM $\rightarrow$ JJ NOM, and NOM $\rightarrow$ NN.

6. Extend your answer to the previous question to cover copula constructions with predicative adjectives, such as *Alex is angry*. The interpretation should be ANGRY(ALEX). You should add a verb phrase production VP $\rightarrow \text{V}_{\text{cop}}$ JJ, and a terminal production $\text{V}_{\text{cop}} \rightarrow$ *is*. Show why your grammar extensions result in the correct interpretation.

7. In Figure 11.6 and Figure 11.7, we treat the plurals *shoots* and *leaves* as entities. Revise $G_2$ so that the interpretation of *Alex eats leaves* is $\forall x.(\text{LEAF}(x) \Rightarrow \text{EATS}(\text{ALEX}, x))$, and show the resulting perceptron update.

8. Statements like *every student eats a pizza* have two possible interpretations, depending on quantifier scope:

$$\forall x \exists y \text{PIZZA}(y) \wedge (\text{STUDENT}(x) \Rightarrow \text{EATS}(x, y)) \quad [11.32]$$
$$\exists y \forall x \text{PIZZA}(y) \wedge (\text{STUDENT}(x) \Rightarrow \text{EATS}(x, y)) \quad [11.33]$$

   Explain why these interpretations really are different, and modify the grammar $G_2$ so that it can produce both interpretations.

9. Derive Equation 11.27.

10. [todo: not sure this works] Download the GeoQuery data, get some deterministic parser that overgenerates, and try to learn a reranker that selects the correct logical form.

11. In the GeoQuery domain, give a natural language query that has multiple plausible semantic interpretations with the same denotation. List both interpretaions and the denotation.

**Hint:** There are many ways to do this, but one approach involves using toponyms (place names) that could plausibly map to several different entities in the model.

# Chapter 12

# Predicate-argument semantics

In this chapter, we consider more "lightweight" semantic representations. These semantic representations discard some aspects of first-order predicate calculus, but focus on predicate-argument structures. Let's start with an example sentence:

(12.1) Asha gives Boyang a book.

The predicate calculus representation of this sentence would be written,

$$\exists x.\text{BOOK}(x) \wedge \text{GIVE}(\texttt{Asha}, \texttt{Boyang}, x) \quad [12.1]$$

In this representation, we define variable $x$ for the book, and we link the strings *Asha* and *Boyang* to entities `Asha` and `Boyang`. Because the action of giving involves a giver, a recipient, and a gift, the predicate GIVE must take three arguments.

Now suppose we have additional information about the event, such as,

(12.2) Yesterday, Asha reluctantly gave Boyang a book.

One possible to solution is to extend the predicate GIVE to take additional arguments,

$$\exists x.\text{BOOK}(x) \wedge \text{GIVE}(\texttt{Asha}, \texttt{Boyang}, x, \texttt{yesterday}, \texttt{reluctantly}) \quad [12.2]$$

But this is clearly unsatisfactory: *yesterday* and *relunctantly* 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 $e$,

$$\exists e, x.\text{GIVE-EVENT}(e) \wedge \text{GIVER}(e, \texttt{Asha}) \wedge \text{GIFT}(e, x) \wedge \text{BOOK}(e, x) \wedge \text{RECIPIENT}(e, \texttt{Boyang}) \\ \wedge \text{TIME}(e, \texttt{Yesterday}) \wedge \text{MANNER}(e, \texttt{reluctantly})$$

In this way, each argument of the event — the giver, the recipient, the gift — can be represented with a relation of its own, linking the argument to the event $e$. The expression $\text{GIVER}(e, \texttt{Asha})$ says that `Asha` plays the **role** of GIVER in the event. This reformulation nicely handles the problem of optional information such as the time or manner of the event, which are called **adjuncts**. Unlike arguments, adjuncts are not a mandatory part of the relation, but under this representation, they can be expressed with additional logical relations that are conjoined to the semantic interpretation of the setnence. [1]

The event semantic representation can be applied to nested clauses, e.g.,

(12.3) Chris sees Asha pay Boyang.

This is done by using the event variable as an argument:

$$\exists e_1, e_2. \text{SEE-EVENT}(e_1) \wedge \text{SEER}(e_1, \texttt{Chris}) \wedge \text{SIGHT}(e_1, e_2) \\ \wedge \text{PAY-EVENT}(e_2) \wedge \text{PAYER}(e_2, \texttt{Asha}) \wedge \text{PAYEE}(e_2, \texttt{Boyang}) \quad [12.3]$$

As with first-order predicate calculus, the goal of event semantics is to provide a representation that generalizes over many surface forms. Consider the following paraphrases of (12.1):

(12.4) Asha gives a book to Boyang.

(12.5) A book is given to Boyang by Asha.

(12.6) A book is given by Asha to Boyang.

(12.7) The gift of a book from Asha to Boyang ...

All have the same event semantic meaning, given in Equation 12.1. Note that the final example does not include a verb! Events are often introduced by verbs, but not always: in this final example, the noun *gift* introduces the same predicate, with the same accompanying arguments.

**Semantic role labeling** (SRL) is a relaxed form of semantic parsing, in which each semantic role is filled by a set of tokens from the text itself. This is sometimes called "shallow semantics" because, unlike model-theoretic semantic parsing, role fillers need not be symbolic expressions with denotations in some world model. A semantic role labeling system is required to identify all predicates, and then specify the spans of text that fill each role, when possible. To get a sense of the task, here is a more complicated example:

[1] This 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).

(12.8) Boyang wants Asha to give him a linguistics book.

In this example, there are two predicates, expressed by the verbs *want* and *give*. Thus, a semantic role labeler should return the following output:

- (PREDICATE : *wants*, WANTER : *Boyang*, DESIRE : *Asha to give him a linguistics book*)
- (PREDICATE : *give*, GIVER : *Asha*, RECIPIENT : *him*, GIFT : *a linguistics book*)

In the example, *Boyang* and *him* may refer to the same person, but this would be ignored in semantic role labeling. However, in other predicate-argument representations, such as **Abstract Meaning Representation (AMR)**, such references must be resolved. We will return to AMR in § 12.3, but first, let us further consider the notion of semantic roles.

## 12.1 Semantic roles

As discussed so far, event semantics requires specifying a number of additional logical relations to link arguments to events: GIVER, RECIPIENT, SEER, SIGHT, etc. Indeed, every predicate requires a set of logical relations to express its own arguments. In contrast, adjuncts such as TIME and MANNER are shared across many types of events. A natural question is whether it is possible to treat mandatory arguments more like adjuncts, by identifying a set of generic argument types that are shared across many event predicates. This can be further motivated by examples involving semantically related verbs:

(12.9) Asha gave Boyang a book.

(12.10) Asha loaned Boyang a book.

(12.11) Asha taught Boyang a lesson.

(12.12) Asha gave Boyang a lesson.

In the first two examples, the roles of Asha, Boyang, and the book are nearly identical. The third example is slightly different, but the fourth example shows that the roles of GIVER and TEACHER can be viewed as related.

One way to think about the relationship between roles such as GIVER and TEACHER is by enumerating the set of properties that an entity typically possesses when it fulfills these roles: givers and teachers are usually animate and "volitional" (meaning that they choose to enter into the action).[2] In contrast, the thing that gets loaned or taught is usually not animate or volitional; furthermore, it is unchanged by the event.

[2]There 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.

| | *Asha* | *gave* | *Boyang* | *a book* |
|---|---|---|---|---|
| **VerbNet** | AGENT | | RECIPIENT | THEME |
| **PropBank** | ARG0: giver | | ARG2: entity given to | ARG1: thing given |
| **FrameNet** | DONOR | | RECIPIENT | THEME |
| | *Asha* | *taught* | *Boyang* | *algebra* |
| **VerbNet** | AGENT | | RECIPIENT | TOPIC |
| **PropBank** | ARG0: teacher | | ARG2: student | ARG1: subject |
| **FrameNet** | TEACHER | | STUDENT | SUBJECT |

Figure 12.1: Example semantic annotations according to VerbNet, PropBank, and FrameNet

Building on these ideas, **thematic roles** generalize across predicates by leveraging the shared semantic properties of typical role fillers (Fillmore, 1968). For example, in examples (12.9-12.12), Asha plays a similar role in all four sentences, which we will call the **agent**. This reflects a number of shared semantic properties: she is the one who is actively and intentionally performing the action, while Boyang is a more passive participant; the book and the lesson would play a different role, as non-animate participants in the event.

Let us now consider a few well-known approaches to semantic roles. Example annotations from each of these systems are shown in Figure 12.1.

### 12.1.1 VerbNet

**VerbNet** (Kipper-Schuler, 2005) is a lexicon of verbs, and it includes thirty "core" thematic roles played by arguments to these verbs. Here are some example roles, accompanied by their definitions from the VerbNet Guidelines.[3]

- AGENT: "ACTOR in an event who initiates and carries out the event intentionally or consciously, and who exists independently of the event."
- PATIENT: "UNDERGOER in an event that experiences a change of state, location or condition, that is causally involved or directly affected by other participants, and exists independently of the event."
- RECIPIENT: "DESTINATION that is animate"
- THEME: "UNDERGOER that is central to an event or state that does not have control over the way the event occurs, is not structurally changed by the event, and/or is characterized as being in a certain position or condition throughout the state."

[3] http://verbs.colorado.edu/verb-index/VerbNet_Guidelines.pdf

- TOPIC: "THEME characterized by information content transferred to another participant."

VerbNet roles are organized in a hierarchy, so that a TOPIC is a type of THEME, which in turn is a type of UNDERGOER, which is a type of PARTICIPANT, the top-level category.

In addition, VerbNet organizes verb senses into a class hierarchy, in which verb senses that have similar meanings are grouped together. Recall from § 3.2 that multiple meanings of the same word are called **senses**, and that WordNet identifies senses for many English words. VerbNet builds on WordNet, so that verb classes are identified by the WordNet senses of the verbs that they contain. For example, the verb class `give-13.1` includes the first WordNet sense of *loan* and the second WordNet sense of *lend*.

Each VerbNet class or subclass takes a set of thematic roles. For example, `give-13.1` takes arguments with the thematic roles of AGENT, THEME, and RECIPIENT;[4] the predicate TEACH takes arguments with the thematic roles AGENT, TOPIC, RECIPIENT, and SOURCE.[5] So according to VerbNet, *Asha* and *Boyang* play the roles of AGENT and RECIPIENT in the sentences,

(12.13) Asha gave Boyang a book.

(12.14) Asha taught Boyang algebra.

The *book* and *algebra* are both THEMES, but *algebra* is a subcategory of THEME — a TOPIC — because it consists of information content that is given to the receiver.

### 12.1.2 Proto-roles and PropBank

Detailed thematic role inventories of the sort used in VerbNet are not universally accepted. For example, (Dowty, 1991, pp. 547) notes that "Linguists have often found it hard to agree on, and to motivate, the location of the boundary between role types." He argues that a solid distinction can be identified between just two **proto-roles**, which have a number of distinguishing characteristics:

- PROTO-AGENT: volitional involvement in the event or state; sentience and/or perception; causing an event or change of state in another participant; movement; exists independently of the event.

[4] `https://verbs.colorado.edu/verb-index/vn/give-13.1.php`
[5] `https://verbs.colorado.edu/verb-index/vn/transfer_mesg-37.1.1.php`

- PROTO-PATIENT: undergoes change of state; causally affected by another participant; stationary relative to the movement of another participant; does not exist independently of the event.[6]

In the examples in Figure 12.1, Asha has most of the proto-agent properties: in giving the book to Boyang, she is acting volitionally (as opposed to *Boyang got a book from Asha*, in which it is not clear whether Asha gave up the book willingly); she is sentient; she causes a change of state in Boyang; she exists independently of the event. Boyang has some proto-agent properties (for example, he is sentient and exists independently of the event), but he also some proto-patient properties: he is the one who is causally affected and who undergoes change of state in both cases. The book that Asha gives Boyang has fewer still of the proto-agent properties — it is not volitional or sentient, and it has no causal role — but it also has few proto-patient properties, as it does not undergo change of state and is not stationary.

The **Proposition Bank**, or PropBank (Palmer et al., 2005), builds on this basic agent-patient distinction, as a middle ground between generic thematic roles and predicate-specific "deep roles." Each verb is linked to a list of numbered arguments, with ARG0 as the proto-agent and ARG1 as the proto-patient. Additional numbered arguments are verb-specific. For example, for the predicate TEACH,[7] the arguments are:

- ARG0: the teacher
- ARG1: the subject
- ARG2: the student(s)

Verbs may have any number of arguments: for example, WANT and GET have five, while EAT has only ARG0 and ARG1. In addition to the semantic arguments found in the frame files, roughly a dozen general-purpose **adjuncts** may be used in combination with any verb. These are shown in Table 12.1.

PropBank-style semantic role labeling is annotated over the entire Penn Treebank. This annotation includes the sense of each verbal predicate, as well as the argument spans.

### 12.1.3 FrameNet

Semantic **frames** are descriptions of situations or events. Frames may be **evoked** by one of their **lexical units** (often a verb, but not always), and they include some number of

[6] Reisinger 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.

[7] http://verbs.colorado.edu/propbank/framesets-english-aliases/teach.html

| | | |
|---|---|---|
| TMP | time | *Boyang ate a bagel* [AM-TMP *yesterday*]. |
| LOC | location | *Asha studies in* [AM-LOC *Stuttgart*] |
| MOD | modal verb | *Asha* [AM-MOD *will*] *study in Stuttgart* |
| ADV | general purpose | [AM-ADV *Luckily*], *Asha knew algebra.* |
| MNR | manner | *Asha ate* [AM-MNR *aggressively*]. |
| DIS | discourse connective | [AM-DIS *However*], *Asha prefers algebra.* |
| PRP | purpose | *Barry studied* [AM-PRP *to pass the bar*]. |
| DIR | direction | *Workers dumped burlap sacks* [AM-DIR *into a bin*]. |
| NEG | negation | *Asha does* [AM-NEG *not*] *know algebra.* |
| EXT | extent | *Prices increased* [AM-EXT *4%*]. |
| CAU | cause | *Asha returned the book* [AM-CAU *because it was overdue*]. |

Table 12.1: PropBank adjuncts (Palmer et al., 2005), sorted by frequency in the corpus

**frame elements**, which are like roles (Fillmore, 1976). For example, the act of teaching is a frame, and can be evoked by the verb *taught*; the associated frame elements include the teacher, the student(s), and the subject being taught. Frame semantics has played a significant role in the history of artificial intelligence, in the work of Minsky (1974) and Schank and Abelson (1977). In natural language processing, the theory of frame semantics has been implemented in **FrameNet** (Fillmore and Baker, 2009), which consists of a lexicon of roughly 1000 frames, and a corpus of more than 200,000 "exemplar sentences," in which the frames and their elements are annotated.[8]

Rather than seeking to link semantic roles such as TEACHER and GIVER into thematic roles such as AGENT, FrameNet aggressively groups verbs into frames, and links semantically-related roles across frames. For example, the following two sentences would be annotated identically in FrameNet:

(12.15) Asha taught Boyang algebra.

(12.16) Boyang learned algebra from Asha.

This is because *teach* and *learn* are both lexical units in the EDUCATION_TEACHING frame. Furthermore, roles can be shared even when the frames are distinct, as in the following two examples:

(12.17) Asha gave Boyang a book.

(12.18) Boyang got a book from Asha.

[8]These statistics are accurate at the time of this writing, in 2017. Current details can be found at the website, `https://framenet.icsi.berkeley.edu/`

The GIVING and GETTING frames both have RECIPIENT and THEME elements, so Boyang and the book would play the same role. Asha's role is different: she is the DONOR in the GIVING frame, and the SOURCE in the GETTING frame. FrameNet makes extensive use of multiple inheritance to share information across frames and frame elements: for example, the COMMERCE_SELL and LENDING frames inherit from GIVING frame.

## 12.2 Semantic role labeling

The task of semantic role labeling is to identify the parts of the sentence comprising the semantic roles. In English, this task is typically performed on the PropBank corpus, with the goal of producing outputs in the following form:

(12.19) [ARG0 Asha] [GIVE.01 gave] [ARG2 Boyang's mom] [ARG1 a book] [AM-TMP yesterday].

Note that a single sentence may have multiple verbs, and therefore a given word may be part of multiple role-fillers:

(12.20) [ARG0 Asha] [WANT.01 wanted] [ARG1 Boyang to give her the book].
Asha wanted [ARG0 Boyang] [GIVE.01 to give] [ARG2 her] [ARG1 the book].

### 12.2.1 Semantic role labeling as classification

PropBank is annotated on the Penn Treebank, and annotators used phrasal constituents (§ 8.2.2) to fill the roles. Therefore SRL can be viewed as the task of assigning to each phrase a label from the set $\mathcal{R} = \{\varnothing, \text{PRED}, \text{ARG0}, \text{ARG1}, \text{ARG2}, \ldots, \text{AM-LOC}, \text{AM-TMP}, \ldots\}$, where $\varnothing$ indicates that the phrase plays no role, and PRED indicates that it is the verbal predicate. If we treat semantic role labeling as a classification problem, we obtain the following functional form:

$$\hat{y}_{(i,j)} = \underset{y}{\text{argmax}}\, \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, y, i, j, \rho, \tau), \quad [12.4]$$

where,

- $(i, j)$ indicates the span of a phrasal constituent $(w_i, w_{i+1}, \ldots, w_{j-1})$;[9]
- $\boldsymbol{w}$ represents the sentence as a sequence of tokens;

[9] PropBank 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 12.2, the subcategorization feature around *taught* would be VP $\rightarrow$ VBD NP PP. |

Table 12.2: Features used in semantic role labeling by Gildea and Jurafsky (2002).

- $\rho$ is the index of the predicate verb in $\boldsymbol{w}$;
- $\tau$ is the structure of the phrasal constituent parse of $\boldsymbol{w}$.

Table 12.2 shows the features used in the seminal paper on FrameNet semantic role labeling by Gildea and Jurafsky (2002). By 2005 there were several systems for PropBank semantic role labeling, and their approaches and feature sets are summarized by Carreras and Màrquez (2005). Typical features include: the phrase type, head word, part-of-speech, boundaries, and neighbors of the proposed argument $\boldsymbol{w}_{i:j}$; the word, lemma, part-of-speech, and voice of the verb $w_\rho$ (active or passive), as well as features relating to its frameset; the distance and path between the verb and the proposed argument. In this way, semantic role labeling systems are high-level "consumers" in the NLP stack, using features produced from lower-level components such as part-of-speech taggers and parsers. More comprehensive and contemporary feature sets are enumerated by Das et al. (2014) and Täckström et al. (2015).

A particularly powerful class of features relate to the **syntactic path** between the argument and the predicate. These features capture the sequence of moves required to get from the argument to the verb by traversing the phrasal constituent parse of the sentence. The idea of these features is to capture syntactic regularities in how various arguments are realized. Syntactic path features are best illustrated by example, using the parse tree

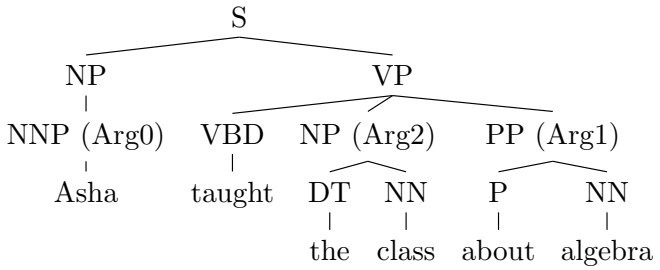



Figure 12.2: Semantic role labeling is often performed on the parse tree for a sentence, labeling individual constituents. [todo: check arg1; show arrows for path features]

in Figure 12.2:

- The path from *Asha* to the verb *taught* is NNP↑NP↑S↓VP↓VBD. The first part of the path, NNP↑NP↑S, means that we must travel up the parse tree from the NNP tag (proper noun) to the S (sentence) constituent. The second part of the path, S↓VP↓VBD, means that we reach the verb by producing a VP (verb phrase) from the S constituent, and then by producing a VBD (past tense verb). This feature is consistent with *Asha* being in subject position, since the path includes the sentence root S.
- The path from *the class* to the verb is NP↑VP↓VBD. This is consistent with *the class* being in object position, since the path passes through the VP node that dominates the verb *taught*.

Because there are many possible path features, it can also be helpful to look at smaller parts: for example, the upward and downward parts can be treated as separate features; another feature might consider whether S appears anywhere in the path.

Rather than using the constituent parse, it is also possible to build features from the **dependency path** between the head word of each argument and the verb (Pradhan et al., 2005). Using the Universal Dependency part-of-speech tagset and dependency relations (Nivre et al., 2016), the dependency path from *Asha* to *taught* is $\text{PROPN} \underset{\text{NSUBJ}}{\leftarrow} \text{VERB}$, because *taught* is the head of a relation of type $\underset{\text{NSUBJ}}{\leftarrow}$ with *Asha*. Similarly, the dependency path from *class* to *taught* is $\text{NOUN} \underset{\text{DOBJ}}{\leftarrow} \text{VERB}$, because *class* heads the noun phrase that is a direct object of *taught*. A more interesting example is *Asha tried to teach the class*, where the path from *Asha* to *tried* is $\text{PROPN} \underset{\text{NSUBJ}}{\leftarrow} \text{VERB} \underset{\text{XCOMP}}{\rightarrow} \text{VERB}$. The right-facing arrow in second relation indicates that *tried* is the head of its XCOMP relation with *teach*.

### 12.2.2 Semantic role labeling as constrained optimization

A potential problem with treating SRL as a classification problem is that there are a number of sentence-level **constraints**, which a classifier might violate.

- For a given verb, there can be only one argument of each type (ARG0, ARG1, etc.)
- Arguments cannot overlap. This problem arises when we are labeling the phrases in a constituent parse tree, as shown in Figure 12.2: if we label the PP *about algebra* as an argument or adjunct, then its children *about* and *algebra* must be labeled as $\varnothing$. The same constraint also applies to the syntactic ancestors of this phrase.

These constraints can be viewed as introducing dependencies across labeling decisions. In structure prediction problems such as sequence labeling and parsing, such dependencies are usually handled by defining additional features over the entire structure, $\boldsymbol{y}$. Efficient inference requires that the global features have a local decomposition that enables dynamic programming: for example, in sequence labeling, the features over $\boldsymbol{y}$ were decomposed into features over pairs of adjacent tags, permitting the application of the Viterbi algorithm for inference. Unfortunately, the constraints that arise in semantic role labeling are less amenable to local decomposition — particularly the constraint that each argument is used only once in the sentence.[10] We therefore consider **constrained optimization** as an alternative solution.

Let the **feasible set** $\mathcal{C}(\tau)$ refer to all labelings that obey the constraints introduced by the parse $\tau$. We can reformulate the semantic role labeling problem as a constrained optimization,

$$\begin{aligned} \max_{\boldsymbol{y}} \quad & \sum_{(i,j)\in\tau} \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, y_{(i,j)}, i, j, \rho, \tau) \\ s.t. \quad & \boldsymbol{y} \in \mathcal{C}(\tau). \end{aligned} \tag{12.5}$$

In this formulation, the objective (shown on the first line) is a separable function of each individual labeling decision, but the constraints (shown on the second line) apply to the overall labeling. The sum $\sum_{(i,j)\in\tau}$ indicates that we are summing over all constituent spans in the parse $\tau$. The expression $s.t.$ in the second line means that we maximize the objective *subject to* the constraints that appear to the right.

#### Integer linear programming

A number of practical algorithms exist for restricted forms of constrained optimization. One such restricted form is **integer linear programming**, in which we optimize a linear

[10] Dynamic programming solutions have been proposed by Tromble and Eisner (2006) and Täckström et al. (2015), but they involves creating a trellis structure whose size is exponential in the number of labels.

objective function over integer variables, with linear constraints. To formulate SRL as an integer linear program, we begin by rewriting the labels as a set of binary variables $\boldsymbol{z} = \{z_{i,j,r}\}$, where,

$$z_{i,j,r} = \begin{cases} 1, & y_{(i,j)} = r \\ 0, & \text{otherwise.} \end{cases} \quad [12.6]$$

Thus, the variables $\boldsymbol{z}$ are a binarized version of the semantic role labeling $\boldsymbol{y}$.

**Objective** Next, we restrict the objective to be a linear function of $\boldsymbol{z}$. We begin with the feature function, $\boldsymbol{f}(\boldsymbol{w}, y_{(i,j)}, i, j, \rho, \tau)$. Such features are typically logical conjunctions involving the label $y_{(i,j)}$. For example:

$$f_j(\boldsymbol{w}, y_{(i,j)}, i, j, \rho, \tau) = \begin{cases} 1, & y_{(i,j)} = \textsc{Arg1} \wedge w_i = \textit{the} \\ 0, & \text{otherwise.} \end{cases} \quad [12.7]$$

This feature is an indicator that takes the value 1 for constituents that are labeled ARG1 and begin with the word *the*. If all features are conjunctions with the label, then the feature function can be rewritten,

$$\boldsymbol{f}(\boldsymbol{w}, y_{(i,j)}, i, j, \rho, \tau) = \sum_{r \in \mathcal{R}} z_{i,j,r} \times \boldsymbol{g}(\boldsymbol{w}, i, j, r, \rho, \tau), \quad [12.8]$$

where $\mathcal{R}$ is the set of all possible labels, $\{\textsc{Arg0}, \textsc{Arg1}, \ldots, \textsc{Am-Loc}, \ldots, \varnothing\}$. With this change in notation, we can now rewrite the objective,

$$\sum_{(i,j) \in \tau} \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, y_{(i,j)}, i, j, \rho, \tau) = \sum_{(i,j) \in \tau} \boldsymbol{\theta} \cdot \left( \sum_{r \in \mathcal{R}} z_{i,j,r} \times \boldsymbol{g}(\boldsymbol{w}, i, j, r, \rho, \tau) \right) \quad [12.9]$$

$$= \sum_{(i,j) \in \tau} \sum_{r \in \mathcal{R}} z_{i,j,r} \left( \boldsymbol{\theta} \cdot \boldsymbol{g}(\boldsymbol{w}, i, j, r, \rho, \tau) \right) \quad [12.10]$$

$$= \sum_{(i,j) \in \tau} \sum_{r \in \mathcal{R}} z_{i,j,r} \psi_{i,j,r}, \quad [12.11]$$

where $\psi_{i,j,r} \triangleq \boldsymbol{\theta} \cdot \boldsymbol{g}(\boldsymbol{w}, i, j, r, \rho, \tau)$. This objective is clearly a linear function of the variables $\boldsymbol{z} = \{z_{i,j,r}\}$.

**Constraints** Integer linear programming permits linear inequality constraints, of the general form $\mathbf{A}\boldsymbol{z} \leq \boldsymbol{b}$, where the parameters $\mathbf{A}$ and $\boldsymbol{b}$ 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,

$$\forall r \neq \varnothing, \quad \sum_{(i,j) \in \tau} z_{i,j,r} \leq 1. \quad [12.12]$$

Recall that $z_{i,j,r} = 1$ if and only if the span $(i, j)$ has label $r$; this constraint says that for each possible label $r \neq \varnothing$, there can be at most one $(i, j)$ such that $z_{i,j,r} = 1$. This constraint can be written in the form $\mathbf{A}\boldsymbol{z} \leq \boldsymbol{b}$, as you will find if you complete the exercises at the end of the chapter.

Now consider the constraint that labels cannot overlap. Let the function $\pi_\tau(i, j) = \{(i', j')\}$ indicate the set of constituents that are ancestors or descendents of $(i, j)$ in the parse $\tau$. For any $(i, j)$ such that $y_{i,j} \neq \varnothing$, the non-overlapping constraint means that all of its ancestors and descendents $(i', j')$ must be labeled as a non-argument, $y_{i',j'} = \varnothing$. We can write this as a set of linear constraints,

$$\forall (i, j) \in \tau, \quad \sum_{r \neq \varnothing} \left( z_{i,j,r} + \sum_{(i',j') \in \pi_\tau(i,j)} z_{i',j',r} \right) \leq 1. \tag{12.13}$$

We can therefore rewrite the semantic role labeling problem as the following integer linear program,

$$\max_{\boldsymbol{z} \in \{0,1\}^{|\tau|}} \quad \sum_{(i,j) \in \tau} \sum_{r \in \mathcal{R}} z_{i,j,r} \psi_{i,j,r} \tag{12.14}$$

$$s.t. \quad \forall r \neq \varnothing, \quad \sum_{(i,j) \in \tau} z_{i,j,r} \leq 1. \tag{12.15}$$

$$\forall (i, j) \in \tau, \quad \sum_{r \neq \varnothing} \left( z_{i,j,r} + \sum_{(i',j') \in \pi_\tau(i,j)} z_{i',j',r} \right) \leq 1. \tag{12.16}$$

The effectiveness of integer linear programming for semantic role labeling was first demonstrated by Punyakanok et al. (2008).

**Learning with constraints** Learning can be performed in the context of constrained optimization using the usual perceptron or large-margin classification updates. Because constrained inference is generally more time-consuming, a key question is whether it is necessary to apply the constraints during learning. Chang et al. (2008) find that better performance can be obtained by learning *without* constraints, and then applying constraints only when using the trained model to predict semantic roles for unseen data.

**How important are the constraints?** Das et al. (2014) find that an unconstrained, classification-based method performs nearly as well as constrained optimization for FrameNet parsing: while it commits many violations of the "no-overlap" constraint, the overall $F_1$ score is less than one point worse than the score at the constrained optimum. Similar results are obtained for PropBank semantic role labeling by Punyakanok et al. (2008). He et al.

(2017) find that constrained inference makes a bigger impact if the constraints are based on manually-labeled "gold" syntactic parses. This implies that errors from the syntactic parser may limit the effectiveness of the constraints. Punyakanok et al. (2008) hedge against parser error by including constituents from several different parsers; any constituent can be selected from any parse, and additional constraints ensure that overlapping constituents are not selected.

**Implementation** Integer linear programming solvers such as `glpk`,[11] `cplex`,[12] and `Gurobi`[13] allow inequality constraints to be expressed directly in the problem definition, rather than in the matrix form $\mathbf{A}\boldsymbol{z} \leq \boldsymbol{b}$. The time complexity of integer linear programming is theoretically exponential in the number of variables $|\boldsymbol{z}|$, but in practice these off-the-shelf solvers obtain good solutions efficiently. Das et al. (2014) report that the `cplex` solver requires 43 seconds to perform inference on the FrameNet test set, which contains 4,458 predicates.

Recent work has shown that many constrained optimization problems in natural language processing can be solved in a highly parallelized fashion, using optimization techniques such as **dual decomposition**, which are capable of exploiting the underlying problem structure (Rush et al., 2010). Das et al. (2014) apply this technique to FrameNet semantic role labeling, obtaining an order-of-magnitude speedup over `cplex`.

### 12.2.3 Neural semantic role labeling

Neural network approaches to SRL have tended to treat it as a sequence labeling task, using a labeling scheme such as the **BIO notation**, which we previously saw in named entity recognition (§ 7.3). In this notation, the first token in a span of type ARG1 is labeled B-ARG1; all remaining tokens in the span are **inside**, and are therefore labeled I-ARG1. Tokens outside any argument are labeled O. For example:

(12.21) *Asha taught Boyang 's mom about algebra*
B-ARG0 PRED B-ARG2 I-ARG2 I-ARG2 B-ARG1 I-ARG1

We now consider two classes of neural networks that can learn to produce such labelings.

**Convolutional neural networks** One of the first applications of **convolutional neural networks** (§ 2.4) to natural language processing was the task of classifying semantic roles. Collobert et al. (2011) treat the task as a classification problem, using information gathered from across the sentence to compute the label for each token. For example, suppose our goal is to tag the role of word $m$ with respect to verb $v$; then for word $n$, we compute the discrete feature vector, $\boldsymbol{f}(\boldsymbol{w}, n, v, m)$, which would include the lower-case word $w_m$, and

[11] https://www.gnu.org/software/glpk/
[12] https://www-01.ibm.com/software/commerce/optimization/cplex-optimizer/
[13] http://www.gurobi.com/

the distances $m - n$ and $m - v$. These features are then used as the inputs to a nonlinear prediction model, as follows:

- Each discrete feature is associated with a dense vector embedding, and these embeddings are concatenated, resulting in a dense vector $\boldsymbol{x}_m^{(0)}$. The horizontal concatenation of the dense embeddings for all words in the sentence is $\mathbf{X}^{(0)} = [\boldsymbol{x}_0^{(0)}, \boldsymbol{x}_1^{(0)}, \ldots, \boldsymbol{x}_M^{(0)}]$.
- Next, a convolutional operation is applied to merge information across words, $\mathbf{X}^{(1)} = \mathbf{C}\mathbf{X}^{(0)}$. Thus, $\boldsymbol{x}_m^{(1)}$ contains information about the word $w_m$, but also about its near neighbors. (Special padding vectors are included on the left and right ends of the matrix $\mathbf{X}^{(0)}$ before convolution.)
- To convert the matrix $\mathbf{X}^{(1)}$ back to a vector $\boldsymbol{z}^{(1)}$, Collobert *et al.* apply **max pooling**. We will write $\boldsymbol{z} = \text{MaxPool}(\mathbf{X})$ to indicate that each $z_j = \max_m(x_{0,j}^{(1)}, x_{1,j}^{(1)}, \ldots, x_{M,j}^{(1)})$.
- The vector $\boldsymbol{z}^{(1)}$ is then passed through several feedforward layers, $\boldsymbol{z}^{(i)} = \boldsymbol{g}(\boldsymbol{\Theta}^{(i)}\boldsymbol{z}^{(i-1)})$, where $\boldsymbol{\Theta}^{(i)}$ is a matrix of weights and $\boldsymbol{g}$ is an elementwise nonlinear transformation.[14]
- At the output layer $\boldsymbol{z}^{(K)}$ is used to make a prediction, $\hat{y} = \text{argmax}_y \boldsymbol{\Theta}^{(y)}\boldsymbol{z}^{(K)}$. [todo: consider making this a figure/algorithm]

This model is very similar to the classifier described in Figure 2.4; the key difference is that the inclusion of distance feature embeddings to compute distinct labels for each word in the span. Collobert et al. (2011) apply this model without regard to constraints, so in principle it could produce labelings that include each argument multiple times. The parameters of the model include the word and feature embeddings that constitute $\mathbf{X}^{(0)}$, the convolution matrix $\mathbf{C}$, the feedforward weight matrices $\boldsymbol{\Theta}^{(i)}$, and the prediction weights $\boldsymbol{\Theta}^{(y)}$. Each of these parameters is estimated by backpropagated stochastic gradient descent (see § 5.3.1). Collobert et al. (2011) emphasize that **multi-task learning** was essential to get good performance: they train the word embeddings not only to accurately predict PropBank labels, but also to assign a high likelihood to a large corpus of unlabeled data. A more contemporary approach would be to use **pre-trained word embeddings**, which have already been trained to predict words in context, thereby avoiding the cost of jointly training across a large unlabeled dataset.

[14] Collobert *et al.* use a piecewise linear **hard tanh** function for nonlinear transformations,

$$g(x) = \begin{cases} -1, & x < -1 \\ x, & -1 \leq x \leq 1 \\ 1, & x > 1. \end{cases} \quad [12.17]$$

Like the **rectified linear unit (ReLU)** (§ 2.2.1), this function is piecewise linear, and has a gradient that is easy to compute.

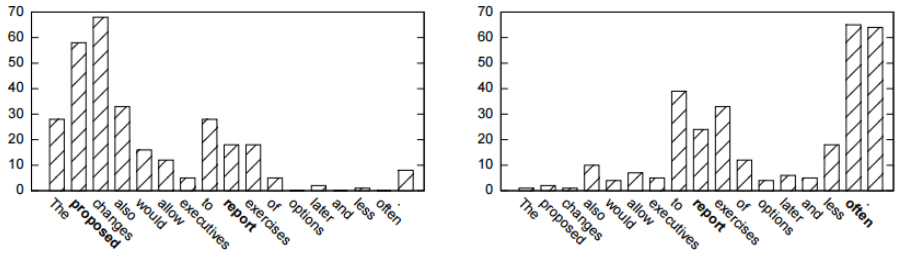



Figure 12.3: Number of features chosen at each word position by the max pooling operation, for tagging the words *proposed* (left) and *often* (right). Figure reprinted from Collobert et al. (2011) [todo: permission]

A key aspect of convolutional neural networks is the use of **pooling** operations, which combine information across a variable-length sequence of vectors into a single vector or matrix. Max pooling is widely used in natural language processing applications, because it enables each element in the vector $\boldsymbol{z}$ to take information from across a sentence or other sequence of text. Figure 12.3 shows the number of "features" taken from each word in a sentence — that is, how often the max operation chooses an element from each word. In each case, the pooling operation emphasizes the word to be tagged, its neighbors, and also the main verb *report*.

**Recurrent neural networks** An alternative neural approach to semantic role labeling is to use **recurrent neural network** models, such as **long short-term memories** (LSTMs; see § 6.5.4 to review how these models are applied to tagging tasks). Zhou and Xu (2015) apply a bidirectional multilayer LSTM 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, $\mathbf{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_M]$. The hidden state of the final LSTM is $\mathbf{Z}^{(K)} = [\boldsymbol{z}_1^{(K)}, \boldsymbol{z}_2^{(K)}, \ldots, \boldsymbol{z}_M^{(K)}]$. The "emission" score for each tag $Y_m = y$ is equal to the inner product $\boldsymbol{\theta}_y \cdot \boldsymbol{z}_m^{(K)}$, and there is also a transition score for each pair of adjacent tags. The complete model can be written,

$$\mathbf{Z}^{(1)} = \text{BiLSTM}(\mathbf{X}) \quad [12.18]$$

$$\mathbf{Z}^{(i)} = \text{BiLSTM}(\mathbf{Z}^{(i-1)}) \quad [12.19]$$

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\text{argmax}} \sum_{m-1}^{M} \boldsymbol{\Theta}^{(y)} \boldsymbol{z}_m^{(K)} + \psi_{y_{m-1}, y_m}. \quad [12.20]$$

Note that the final step maximizes over the entire labeling $\boldsymbol{y}$, and includes a score for each tag transition $\psi_{y_{m-1}, y_m}$. This combination of LSTM and pairwise potentials on tags is an example of an **LSTM-CRF**. The maximization over $\boldsymbol{y}$ is performed by the Viterbi algorithm.

```
(w / want-01
  :ARG0 (b / boy)
  :ARG1 (g / go-02
           :ARG0 b))
```

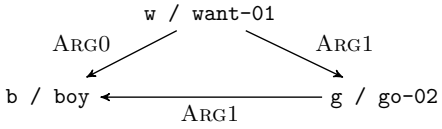



Figure 12.4: Two views of the AMR representation for the sentence *The boy wants to go.*

This model strongly outperformed alternative approaches at the time, including constrained decoding and convolutional neural networks. More recent work has combined recurrent neural network models with constrained decoding, using the $A^*$ search algorithm to search over labelings that are feasible with respect to the constraints (He et al., 2017). This yields small improvements over the method of Zhou and Xu (2015). He et al. (2017) obtain larger improvements by creating an **ensemble** of SRL systems, each trained on an 80% subsample of the corpus. The average prediction across this ensemble is more robust than any individual model.

## 12.3 Abstract Meaning Representation

Semantic role labeling transforms the task of semantic parsing to a labeling task. Consider the sentence,

(12.22) The boy wants to go.

The PropBank semantic role labeling analysis is:

- (PREDICATE : *wants*, ARG0 : *the boy*, ARG1 : *to go*)
- (PREDICATE : *go*, ARG1 : *the boy*)

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 12.4. 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*. Relations are introduced with colons: for example, `:arg0 (b / boy)` indicates a relation of type `arg0` with the newly-introduced variable `b`. Variables can be reused, so that when the variable `b` appears again as an argument to `g`, it is understood to refer to the same boy in both cases. This arrangement is indicated compactly in the graph structure on the right, with edges indicating concepts.

AMR differs from PropBank-style semantic role labeling in a few key ways. First, it reifies each entity as a variable: for example, the *boy* in (12.22) is reified in the variable `b`, which is reused as ARG0 in its relationship with `w / want-01`, and as ARG1 in its relationship with `g / go-02`. Reifying entities as variables also makes it possible to represent the substructure of noun phrases more explicitly. For example, *Asha borrowed the algebra book* would be represented as:

```
(b / borrow-01
   :ARG0 (p / person
            :name (n / name
                     :op1 "Asha"))
   :ARG1 (b2 / book
            :topic (a / algebra)))
```

This indicates that the variable `p` is a person, whose `name` is the variable `n`; that name has one token, the string *Asha*. Similarly, the variable `b2` is a book, and the `topic` of `b2` is a variable `a` whose type is `algebra`. The relations `name` and `topic` are examples of **non-core roles**, which are similar to adjunct modifiers in PropBank. However, AMR's inventory is more extensive, including more than 70 non-core roles, such as negation, time, manner, frequency, and location. Lists and sequences — such as the list of tokens in a name — are described using the roles `op1`, `op2`, etc.

Another key feature of AMR is that a semantic predicate can be introduced by any syntactic element. Consider the following examples, from Banarescu et al. (2013):

(12.23) The boy destroyed the room.

(12.24) the destruction of the room by the boy ...

(12.25) the boy's destruction of the room ...

All these examples have the same semantics in AMR,

```
(d / destroy-01
   :arg0 (b / boy)
   :arg1 (r / room))
```

Note that the noun *destruction* is linked to the verb *destroy*, which is captured by the PropBank frame `destroy-01`. This can happen with adjectives as well: in the phrase *the attractive spy*, the adjective *attractive* is linked to the PropBank frame `attract-01`:

```
(s / spy
   :arg0-of (a / attract-01))
```

In this example, `arg0-of` is an **inverse relation**, indicating that `s` is the `arg0` of the predicate `a`. Inverse relations make it possible for all AMR parses to have a single root concept, which should be the **focus** of the utterance.

There are a number of other important linguistic issues in the design of AMR, which are summarized in the original paper (Banarescu et al., 2013) and the tutorial slides by Schneider et al. (2015). While AMR goes farther than semantic role labeling, it does not link semantically-related frames such as `buy/sell` (as FrameNet does), does not handle quantification (as first-order predicate calculus does), and makes no attempt to handle noun number and verb tense (as PropBank does). A recent survey by Abend and Rappoport (2017) situates AMR with respect to several other semantic representation schemes.

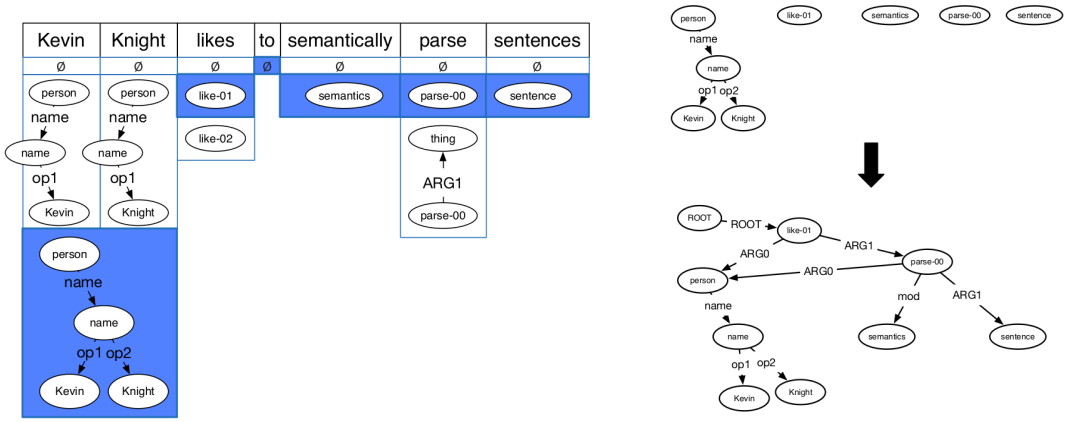
### 12.3.1 AMR Parsing

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.

**Graph-based parsing** One family of approaches to AMR parsing is similar to the graph-based methods that we encountered in syntactic dependency parsing (chapter 10). For these systems (Flanigan et al., 2014), parsing is a two-step process:

1. **Concept identification** (Figure 12.5a). This involves constructing concept subgraphs for individual words or spans of adjacent words. For example, in the sentence, *Asha likes algebra*, we would hope to identify the minimal subtree including just the concept `like-01` for the word *like*, and the subtree `(p / person :name (n / name :op1 Asha))` for the word *Asha*.




(a) Concept identification (b) Relation identification

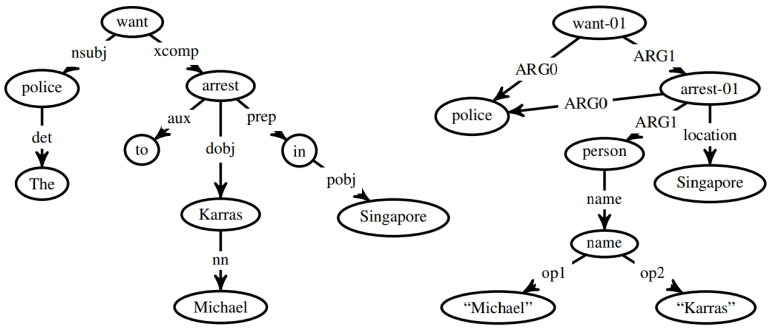
Figure 12.5: Subtasks for Abstract Meaning Representation parsing, from Schneider et al. (2015). [todo: ask for permission, or remake]

2. **Relation identification** (Figure 12.5b). This involves building a directed graph over the concepts, where the edges are labeled by the relation type. AMR imposes a number of constraints on the graph: all concepts must be included, the graph must be **connected** (there must be a path between every pair of nodes in the undirected version of the graph), and every node must have at most one outgoing edge of each type.

Both of these problems are solved by structure prediction. Concept identification requires simultaneously segmenting the text into spans, and labeling each span with a graph fragment containing one or more concepts. This is done by computing a set of features for each candidate span $\boldsymbol{s}$ and concept labeling $\boldsymbol{c}$, and then returning the labeling with the highest overall score.

Relation identification can be formulated as search for the maximum spanning subgraph, under a set of constraints. Each labeled edge has a score, which is computed from features of the concepts. We then search for the set of labeled edges that maximizes the sum of these scores, under the constraint that the resulting graph is well-formed AMR. § 12.2.2 described how constrained optimization can be used for semantic role labeling; similar techniques have been applied to AMR relation identification (Flanigan et al., 2014).

**Transition-based parsing** In many cases, AMR parses are structurally similar to syntactic dependency parses. Figure 12.6 shows one such example. This motivates an alternative approach to AMR parsing: simply modify the syntactic dependency parse until it looks like a good AMR parse. Wang et al. (2015) propose a transition-based method, based on




(a) Dependency tree (b) AMR graph

Figure 12.6: Syntactic dependency parse and AMR graph for the sentence *The police want to arrest Michael Karras in Singapore* (borrowed from Wang et al. (2015)) [todo: permission]

incremental modifications to the syntactic dependency tree (you may review transition-based dependency parsing in § 10.3). 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. They train their system as a classifier, learning to choose the action as would be given by an **oracle** that is capable of reproducing the ground-truth parse. The 2016 SemEval evaluation compared a number of contemporary AMR parsing systems (May, 2016).

## 12.4 Applications of Predicate-Argument Semantics

**Question answering** **Factoid questions** have answers that are single words or phrases, such as *who discovered priors?*, *where was Barack Obama born?*, and *in what year did the Knicks last win the championship?* Shen and Lapata (2007) show that semantic role labeling can be used to answer such questions, by linking them to sentences in a corpus of text. They perform FrameNet semantic role labeling, making heavy use of dependency path features. For each sentence, they produce a weighted **bipartite graph**[15] between FrameNet semantic roles and the words and phrases in the sentence. This is done by first scoring all pairs of semantic roles and assignments, as shown in the top half of Figure 12.8. They then find the bipartite edge cover, which is the minimum weighted subset of edges such that each vertex has at least one edge, as shown in the bottom half of Figure 12.8. After analyzing the question in this manner, Shen and Lapata then find semantically-compatible sentences in the corpus, by performing graph matching on the bipartite graphs for the question and

[15] A bipartite graph is one in which the vertices can be divided into two disjoint sets, and every edge connect a vertex in one set to a vertex in the other.

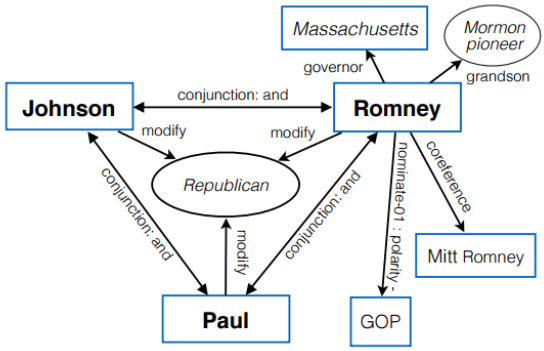



Figure 12.7: Fragment of AMR knowledge network for entity linking. Figure reprinted from Pan et al. (2015) [todo: permission]

candidate answer sentences. Finally, the **expected answer phrase** in the question — typically the *wh*-word — is linked to a phrase in the candidate answer source, and that phrase is returned as the answer.

**Relation extraction** The task of **relation extraction** involves identifying pairs of entities for which a given semantic relation holds. For example, we might like to find all $\langle i, j \rangle$ such that $i$ is the INVENTOR-OF $j$. PropBank semantic role labeling can be applied to this task by identifying sentences whose verb signals the desired relation, and then extracting ARG1 and ARG2 as arguments. (To fully solve this task, these arguments must then be linked to entities, as described in chapter 16.) Christensen et al. (2010) compare the UIUC semantic role labeling system (which uses integer linear programming) against a simpler approach based on surface patterns (Banko et al., 2007). They find that the SRL system is considerably more accurate, but that it is several orders of magnitude slower. Conversely, Barnickel et al. (2009) apply SENNA, a convolutional neural network SRL system (Collobert and Weston, 2008) to the task of identifying biomedical relations (e.g., which genes inhibit or activate each other). In comparison with a strong baseline that applies a set of rules to syntactic dependency structures (Fundel et al., 2007), the SRL system is faster but less accurate. One possible explanation for these divergent results is that the Fundel et al. compare against a baseline which is carefully tuned for performance in a relatively narrow domain, while the system of Banko et al. is designed to analyze text across the entire web.

**Entity linking** Another core task in information extraction is to link mentions of entities (e.g., *Republican candidates like Romney, Paul, and Johnson ...*) to entities in a knowledge base (e.g., `Lyndon Johnson` or `Gary Johnson`). This is often done by examining nearby "collaborator" mentions — in this case, *Romney* and *Paul*. By jointly linking all

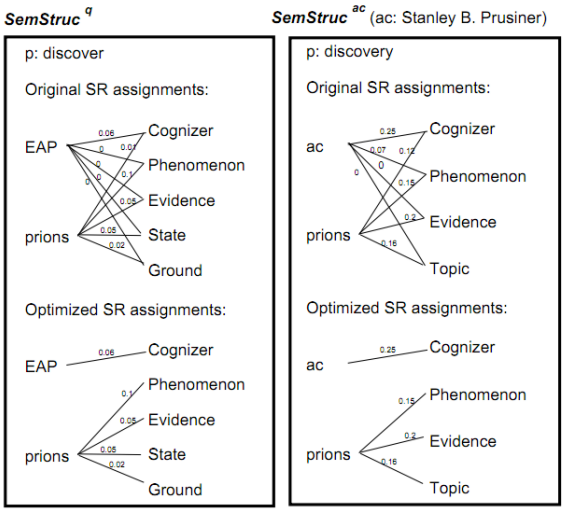



Figure 12.8: FrameNet semantic role labeling is used in factoid question answering, by aligning the semantic roles in the question (q) against those of sentences containing answer candidates (ac). "EAP" is the expected answer phrase, replacing the word *who* in the question. Figure reprinted from Shen and Lapata (2007) [todo: permission]

such mentions, it is possible to arrive at a good overall solution. Pan et al. (2015) apply AMR to this problem. For each entity, they construct a knowledge network based on its semantic relations with other mentions within the same sentence. They then rerank a set of candidate entities, based on the overlap between the entity's knowledge network and the semantic relations present in the sentence (Figure 12.7). When applied to manually labeled AMR annotations, this approach is superior to state-of-the-art supervised methods that have access to labeled examples of linked mentions. Pan et al. also show that the method performs well from automated AMR, and that an AMR-based approach far outperforms a similar method based on PropBank semantic role labeling.[todo: rework for clarity]

## Exercises

1. Write out an event semantic representation for the following sentences. You may make up your own predicates.

   (12.26) *Abigail shares with Max.*

   (12.27) *Abigail reluctantly shares a toy with Max.*

   (12.28) *Abigail hates to share with Max.*

2. Find the PropBank framesets for *share* and *hate* at `http://verbs.colorado.edu/propbank/framesets-english-aliases/`, and rewrite your answers from the previous question, using the thematic roles ARG0, ARG1, and ARG2.

3. Compute the syntactic path features for Abigail and Max in each of the example sentences (12.26) and (12.28) in Question 1, with respect to the verb *share*. If you're not sure about the parse, you can try an online parser such as `http://nlp.stanford.edu:8080/parser/`.

4. Compute the dependency path features for Abigail and Max in each of the example sentences (12.26) and (12.28) in Question 1, with respect to the verb *share*. Again, if 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* and *Max* is OBL according to the Universal Dependency treebank (version 2).

5. PropBank semantic role labeling includes **reference arguments**, such as,

   (12.29) [AM-LOC The bed] on [R-AM-LOC which] I slept broke.[16]

   The label R-AM-LOC indicates that word *which* is a reference to *The bed*, which expresses the location of the event. Reference arguments must have referents: the tag R-AM-LOC can appear only when AM-LOC also appears in the sentence. Show how to express this as a linear constraint, specifically for the tag R-AM-LOC. Be sure to correctly handle the case in which neither AM-LOC nor R-AM-LOC appear in the sentence.

6. Explain how to express the constraints on semantic role labeling in Equation 12.12 and Equation 12.13 in the general form $\mathbf{A}\boldsymbol{z} \geq \boldsymbol{b}$.

7. Download the FrameNet sample data (`https://framenet.icsi.berkeley.edu/fndrupal/fulltextIndex`), and train a bag-of-words classifier to predict the frame that is evoked by each verb in each example. Your classifier should build a bag-of-words from the sentence in which the frame-evoking lexical unit appears. [todo: Somehow limit to one or a few lexical units.] [todo: use NLTK if possible]

8. Download the PropBank sample data, using NLTK (`http://www.nltk.org/howto/propbank.html`). Use a deep learning toolkit such as PyTorch or DyNet to train an LSTM to predict tags. You will have to convert the downloaded instances to a BIO sequence labeling representation first.

9. Produce the AMR annotations for the following examples:

[16] Example from 2013 NAACL tutorial slides by Shumin Wu

(12.30) The girl likes the boy.

(12.31) The girl was liked by the boy.

(12.32) Abigail likes Maxwell Aristotle.

(12.33) The spy likes the attractive boy.

(12.34) The girl doesn't like the boy.

(12.35) The girl likes her dog.

For (12.32), recall that multi-token names are created using `op1`, `op2`, etc. You will need to consult Banarescu et al. (2013) for (12.34), and Schneider et al. (2015) for (12.35). You may assume that *her* refers to *the girl* in this example.

10. Using an off-the-shelf PropBank SRL system,[17] build a simplified question answering system in the style of Shen and Lapata (2007). Specifically, your system should do the following:

    - For each document in a collection, it should apply the semantic role labeler, and should store the output as a tuple.
    - For a question, your system should again apply the semantic role labeler. If any of the roles are filled by a *wh*-pronoun, you should mark that role as the expected answer phrase (EAP).
    - To answer the question, search for a stored tuple which matches the question as well as possible (same predicate, no incompatible semantic roles, and as many matching roles as possible). Align the EAP against its role filler in the stored tuple, and return this as the answer.

    To evaluate your system, download a set of three news articles on the same topic, and write down five factoid questions that should be answerable from the articles. See if your system can answer these questions correctly. (If this problem is assigned to an entire class, you can build a large-scale test set and compare various approaches.)

[17] At the time of writing, the following systems are availabe: 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/`).

# Chapter 13

# Distributional and distributed semantics

A recurring theme in natural language processing is the complexity of the mapping from words to meaning. In chapter 3, we saw that a single word form, like *bank*, can have multiple meanings; conversely, a single meaning may be created by multiple surface forms, a lexical semantic relationship known as **synonymy**. Despite this complex mapping between words and meaning, natural language processing systems usually rely on words as the basic unit of analysis. This is especially true in semantics: the logical and frame semantic methods from the previous two chapters rely on hand-crafted lexicons that map from words to semantic predicates. But how can we analyze texts that contain words that we haven't seen before? This chapter describes methods that learn representations of word meaning by analyzing unlabeled data, vastly improving the generalizability of natural language processing systems. The theory that makes it possible to acquire meaningful representations from unlabeled data is the **distributional hypothesis**.

## 13.1 The distributional hypothesis

Here's a word you may not know: *tezgüino*.[1] If you do not know the meaning of *tezgüino*, then you are in the same situation as a natural language processing system when it encounters a word that did not appear in its training data. Now suppose you see that *tezgüino* is used in the following contexts:

- **C1**: *A bottle of _____ is on the table*
- **C2**: *Everybody likes _____.*
- **C3**: *Don't have _____ before you drive.*

[1] The example is from Lin (1998).

| | C1 | C2 | C3 | C4 | ... |
|---|---|---|---|---|---|
| *tezgüino* | 1 | 1 | 1 | 1 | |
| *loud* | 0 | 0 | 0 | 0 | |
| *motor oil* | 1 | 0 | 0 | 1 | |
| *tortillas* | 0 | 1 | 0 | 1 | |
| *choices* | 0 | 1 | 0 | 0 | |
| *wine* | 1 | 1 | 1 | 0 | |

Table 13.1: Distributional statistics for *tezgüino* and five related terms

- **C4**: *We make _____ out of corn.*

What other words fit into these contexts? How about: *loud, motor oil, tortillas, choices, wine*? Each row of Table 13.1 is a vector that summarizes the contextual properties for each word, with a value of one for contexts in which the word can appear, and a value of zero for contexts in which it cannot. Based on these vectors, we can conclude:

- *wine* is very similar to *tezgüino*;
- *motor oil* and *tortillas* are fairly similar to *tezgüino*;
- *loud* is different.

These vectors, which we will call **word representations**, describe the **distributional** properties of each word. Does vector similarity imply semantic similarity? This is the **distributional hypothesis**, stated by Firth (1957) as: "You shall know a word by the company it keeps." The distributional hypothesis has stood the test of time: distributional statistics are a core part of language technology today, because they make it possible to leverage large amounts of unlabeled data to learn about rare words that do not appear in labeled training data.

Distributional statistics have the striking ability to capture lexical semantic relationships such as analogies. Figure 13.1 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

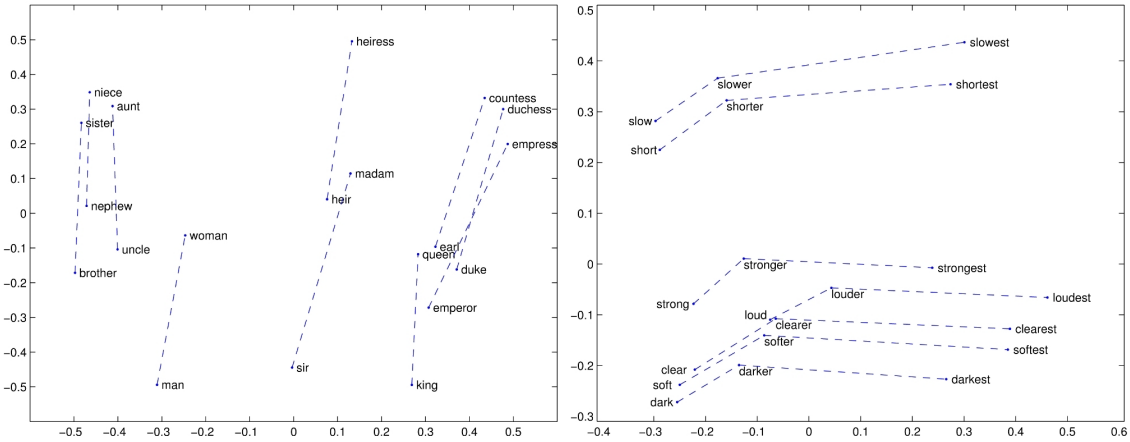



Figure 13.1: Lexical semantic relationships have regular linear structures in two dimensional projections of distributional statistics. From `http://nlp.stanford.edu/projects/glove/`.[todo: redo to make words bigger?]

chapter. However, distributed representations can also be learned in a supervised fashion from labeled data, as in the neural classification models encountered in chapter 2.

## 13.2 Design decisions for word representations

There are many approaches for computing word representations, but most can be distinguished on three main dimensions of variation: the nature of the representation, the source of contextual information, and the estimation procedure.

### 13.2.1 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).

Dense vector word embeddings are well suited for neural network architectures, 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** (§ 13.4), 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, or whether to estimate distinct embeddings for each word sense. 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. (2016) 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.

### 13.2.2 Context

The distributional hypothesis says that word meaning is related to the "contexts" in which the word appears, but context can be defined in many ways. In the *tezgüino* example, contexts are entire sentences, but in practice there are far too many sentences for this to work. At the opposite extreme, the context could be defined as the immediately preceding word; this is the context considered in Brown clusters. WORD2VEC takes an intermediate approach, using local neighborhoods of words (e.g., $h = 5$) as contexts (Mikolov et al., 2013). Contexts can also be much larger: for example, in **latent semantic analysis**, each word's context vector includes an entry per document, with a value of one if the word appears in the document (Deerwester et al., 1990); in **explicit semantic analysis**, these documents are Wikipedia pages (Gabrilovich and Markovitch, 2007).

Words in context can be labeled by their position with respect to the target word $w_m$ (e.g., two words before, one word after), which makes the resulting word representations more sensitive to syntactic differences (Ling et al., 2015). Another way to incorporate syntax is to perform parsing as a preprocessing step, and then form context vectors from the set of dependency edges (Levy and Goldberg, 2014) or predicate-argument relations (Lin, 1998). The resulting context vectors for several of these methods are shown in Table 13.2.

The choice of context has a profound effect on the resulting representations, which can be viewed in terms of word similarity. Applying latent semantic analysis (§ 13.3) to contexts of size $h = 2$ and $h = 30$ yields the following nearest-neighbors for the word *dog*:[2]

- $(h = 2)$: *cat, horse, fox, pet, rabbit, pig, animal, mongrel, sheep, pigeon*
- $(h = 30)$: *kennel, puppy, pet, bitch, terrier, rottweiler, canine, cat, to bark, Alsatian*

Which word list is better? Each word in the $h = 2$ list is an animal, reflecting the fact that locally, the word *dog* tends to appear in the same contexts as other animal types (e.g., *pet*

[2]The 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/`

| *The moment one **learns** English, complications set in.* | |
|---|---|
| Brown Clusters (Brown et al., 1992) | $\{$*one*$\}$ |
| WORD2VEC (Mikolov et al., 2013) $(h = 2)$ | $\{$*moment, one, English, complications*$\}$ |
| Structured WORD2VEC (Ling et al., 2015) $(h = 2)$ | $\{(\textit{moment}, -2), (\textit{one}, -1), (\textit{English}, +1), (\textit{complications}, +2)\}$ |
| Dependency contexts (Levy and Goldberg, 2014) | $\{(\textit{one}, \text{NSUBJ}), (\textit{English}, \text{DOBJ}), (\textit{moment}, \text{ACL}^{-1})\}$ |

Table 13.2: Contexts for the word *learns*, according to various word representations. For dependency context, $(\textit{one}, \text{NSUBJ})$ means that there is a relation of type NSUBJ (nominal subject) **to** the word *one*, and $(\textit{moment}, \text{ACL}^{-1})$ means that there is a relation of type ACL (adjectival clause) **from** the word *moment*.

*the dog*, *feed the dog*). In the $h = 30$ list, nearly everything is dog-related, including specific breeds such as *rottweiler* and *Alsatian*. The list also includes words that are not animals (*kennel*), and in one case (*to bark*), is not a noun at all. The 2-word context window is more sensitive to syntax, while the 30-word window is more sensitive to topic.

### 13.2.3 Estimation

Word embeddings are estimated by optimizing some objective: the likelihood of a set of unlabeled data (or a closely related quantity), or the reconstruction of a matrix of context counts, similar to Table 13.1.

**Maximum likelihood estimation** Likelihood-based optimization is derived from the objective $\log \text{p}(\boldsymbol{w}; \boldsymbol{u})$, where $\boldsymbol{u}_i$ is the embedding of word $i$, and $\boldsymbol{w} = \{w_m\}_{m=1}^{M}$ is a corpus, represented as a list of $M$ tokens. Recurrent neural network language models (§ 5.3) optimize this objective directly, backpropagating to the input word embeddings through the recurrent structure. However, state-of-the-art word embeddings employ huge corpora with hundreds of billions of tokens, and recurrent architectures are difficult to scale to such data. As a result, likelihood-based word embeddings are usually 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 $\boldsymbol{u}_i$ and a context embedding $\boldsymbol{v}_j$. These embeddings can be obtained by minimizing the norm of

the reconstruction error,

$$\min_{\boldsymbol{u},\boldsymbol{v}} ||\mathbf{C} - \tilde{\mathbf{C}}(\boldsymbol{u}, \boldsymbol{v})||_F, \tag{13.1}$$

where $\tilde{\mathbf{C}}(\boldsymbol{u}, \boldsymbol{v})$ is the approximate reconstruction resulting from the embeddings $\boldsymbol{u}$ and $\boldsymbol{v}$, and $||\mathbf{X}||_F$ indicates the Frobenius norm, $\sum_{i,j} x_{i,j}^2$. Rather than factoring the matrix of word-context counts, it is useful to transform these counts using information-theoretic metrics such as **pointwise mutual information** (PMI), described in the next section.

## 13.3 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),

$$\min_{\mathbf{U}\in\mathbb{R}^{V\times K},\mathbf{S}\in\mathbb{R}^{K\times K},\mathbf{V}\in\mathbb{R}^{|\mathcal{C}|\times K}} ||\mathbf{C} - \mathbf{U}\mathbf{S}\mathbf{V}^\top||_F \tag{13.2}$$

$$s.t. \quad \mathbf{U}^\top\mathbf{U} = \mathbb{I} \tag{13.3}$$

$$\mathbf{V}^\top\mathbf{V} = \mathbb{I} \tag{13.4}$$

$$\forall i \neq j, \mathbf{S}_{i,j} = 0, \tag{13.5}$$

where $V$ is the size of the vocabulary and $|\mathcal{C}|$ is the number of contexts. The word embeddings can then be set to the rows of the matrix $\mathbf{U}$. The matrix $\mathbf{S}$ is constrained to be diagonal (these are called the singular values), and the columns of the product $\mathbf{S}\mathbf{V}^\top$ provide descriptions of the contexts. Each element $c_{i,j}$ is then reconstructed as a **bilinear product**,

$$c_{i,j} \approx \sum_{k=1}^{K} u_{i,k} s_k v_{j,k}, \tag{13.6}$$

and 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`.[3]

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;

[3] An 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 $V \times |\mathcal{C}|$.

Church and Hanks, 1990), which captures the degree of association between word $i$ and context $j$,

$$\text{PMI}(i,j) = \log \frac{\text{p}(i,j)}{\text{p}(i)\text{p}(j)} = \log \frac{\text{p}(i \mid j)\text{p}(j)}{\text{p}(i)\text{p}(j)} = \log \frac{\text{p}(i \mid j)}{\text{p}(i)} \quad [13.7]$$

$$= \log \text{count}(i,j) - \log \sum_{i'=1}^{V} \text{count}(i',j) - \log \sum_{j' \in \mathcal{C}} \text{count}(i,j') + \log \sum_{i'=1}^{V} \sum_{j' \in \mathcal{C}} \text{count}(i',j'), \quad [13.8]$$

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 contexts. When word $i$ is statistically associated with context $j$, the ratio will be greater than one, so $\text{PMI}(i,j) > 0$. The PMI transformation focuses latent semantic analysis on reconstructing strong word-context associations, rather than on reconstructing large counts.

The PMI is negative when a word and context occur together less often than if they were independent, but such negative correlations are unreliable because counts of rare events have high variance. Furthermore, the PMI is undefined when $\text{count}(i,j) = 0$. **Positive PMI** (PPMI) is defined as,

$$\text{PPMI}(i,j) = \begin{cases} \text{PMI}(i,j), & \text{p}(i \mid j) > \text{p}(i) \\ 0, & \text{otherwise.} \end{cases} \quad [13.9]$$

Bullinaria and Levy (2007) compare a range of matrix transformations for latent semantic analysis, using a battery of tasks related to word meaning and word similarity (for more on evaluation, see § 13.6). They find that PPMI-based latent semantic analysis yields strong performance on a battery of tasks related to word meaning: for example, PPMI-based LSA vectors can be used to solve multiple-choice word similarity questions from the Test of English as a Foreign Language (TOEFL), obtaining 85% accuracy.

## 13.4 Brown clusters

Learning algorithms like perceptron and conditional random fields often perform better with discrete feature vectors. A simple way to obtain discrete representations from distributional statistics is by clustering (§ 4.1.1), so that words in the same cluster have similar distributional statistics. This can help in downstream tasks, by sharing features between all words in the same cluster. However, there is an obvious tradeoff: if the number of clusters is too small, the words in each cluster will not have much in common; if the number of clusters is too large, then the learner will not see enough examples from each cluster to generalize.

The solution to this problem is **hierarchical clustering**: using the distributional statistics to induce a tree-structured representation. Fragments of Brown cluster trees are shown

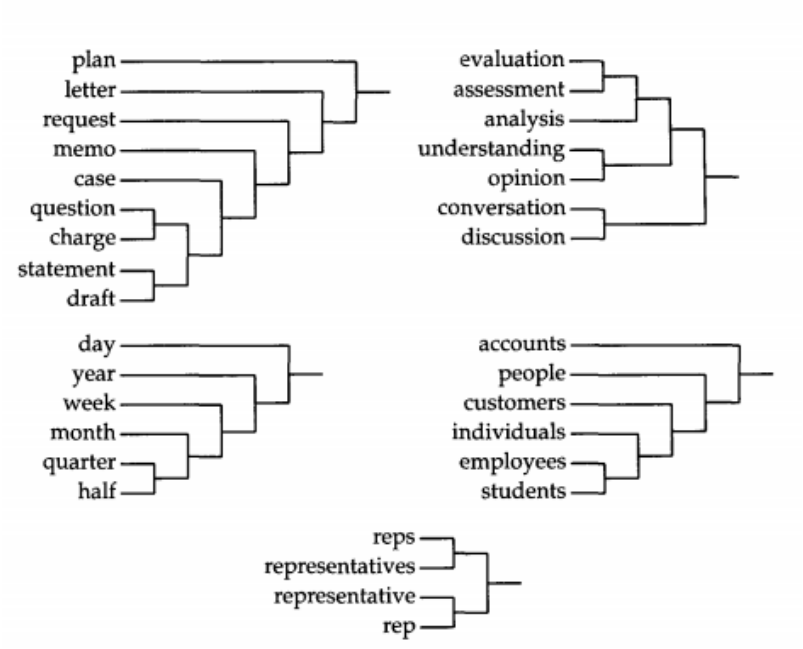



Figure 13.2: Some subtrees produced by bottom-up Brown clustering (Miller et al., 2004) on news text [todo: permission]

in Figure 13.2 and Table 13.3. Each word's representation consists of a binary string describing a path through the tree: `0` for taking the left branch, and `1` for taking the right branch. In the subtree in the upper right of the figure, the representation of the word *conversation* is `01`; the representation of the word *assessment* is `1110`. Bitstring prefixes capture similarity at varying levels of specificity, and it is common to use the first eight, twelve, sixteen, and twenty bits as features in tasks such as named entity recognition (Miller et al., 2004) and dependency parsing (Koo et al., 2008).

Hierarchical trees can be induced from a likelihood-based objective, using a discrete latent variable $k_i \in \{1, 2, \ldots, K\}$ to represent the cluster of word $i$:

$$\log \mathrm{p}(\boldsymbol{w}; \boldsymbol{k}) \approx \sum_{m=1}^{M} \log \mathrm{p}(w_m \mid w_{m-1}; \boldsymbol{k}) \quad [13.10]$$

$$\triangleq \sum_{m=1}^{M} \log \mathrm{p}(w_m \mid k_{w_m}) + \log \mathrm{p}(k_{w_m} \mid k_{w_{m-1}}). \quad [13.11]$$

This is similar to a hidden Markov model, with the crucial difference that each word can be emitted from only a single cluster: $\forall k \neq k_{w_m}, \mathrm{p}(w_m \mid k) = 0$.

| bitstring | ten most frequent words |
|---|---|
| 01111010**0111** | *excited thankful grateful stoked pumped anxious hyped psyched exited geeked* |
| 01111010**100** | *talking talkin complaining talkn bitching tlkn tlkin bragging raving +k* |
| 01111010**1010** | *thinking thinkin dreaming worrying thinkn speakin reminiscing dreamin daydreaming fantasizing* |
| 01111010**1011** | *saying sayin suggesting stating sayn jokin talmbout implying insisting 5'2* |
| 01111010**1100** | *wonder dunno wondered duno donno dno dono wonda wounder dunnoe* |
| 01111010**1101** | *wondering wonders debating deciding pondering unsure wonderin debatin woundering wondern* |
| 01111010**1110** | *sure suree suuure suure sure- surre sures shuree* |

Table 13.3: 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.

Using the objective in Equation 13.11, 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 13.11 is maximized. Although the objective seems to involve a sum over the entire corpus, the score for each merger can be computed from the co-occurrence statistics count$(i, j)$, which counts the number of times a word from cluster $i$ follows a word from cluster $j$. 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**,

$$I(\boldsymbol{k}) = \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} \text{p}(k_1, k_2) \times \text{PMI}(k_1, k_2) \quad [13.12]$$

$$\text{p}(k_1, k_2) = \frac{\text{count}(k_1, k_2)}{\sum_{k_{1'}=1}^{K} \sum_{k_{2'}=1}^{K} \text{count}(k_{1'}, k_{2'})},$$

where $\text{p}(k_1, k_2)$ is the joint probability of a bigram involving a word in cluster $k_1$ followed by a word in $k_2$. This probability and the PMI are both computed from the co-occurrence

**Algorithm 16** Exchange clustering algorithm

**procedure** EXCHANGECLUSTERING($\{\text{count}(\cdot,\cdot)\}, K$)
  **for** $i \in \mathcal{V}$ **do** ▷ Compute unigram counts
    $\text{count}(i) \leftarrow \sum_{j \in \mathcal{C}} \text{count}(i, j)$
  $\{w_1, w_2, \ldots, w_V\} \leftarrow$ Sort-Descending($\{\text{count}(\cdot)\}$)
  **for** $i \in 1 \ldots K$ **do** ▷ Put each of the most common words in its own cluster
    $k_i \leftarrow i$
  **for** $i \in (K+1) \ldots V$ **do** ▷ Iteratively add each word to the clustering
    $k_i \leftarrow K + 1$
    Let $\langle k, k' \rangle$ be the two clusters whose merger maximizes $I(\boldsymbol{k})$ (Equation 13.12)
    Renumber clusters so that $k$ and $k'$ are merged
    Update cluster co-occurrence counts
  **repeat** ▷ Merge the remaining clusters into a tree
    Merge clusters $\langle k, k' \rangle$ to maximize $I(\boldsymbol{k})$.
  **until** only one cluster remains
  **return** Tree structure induced by merge history

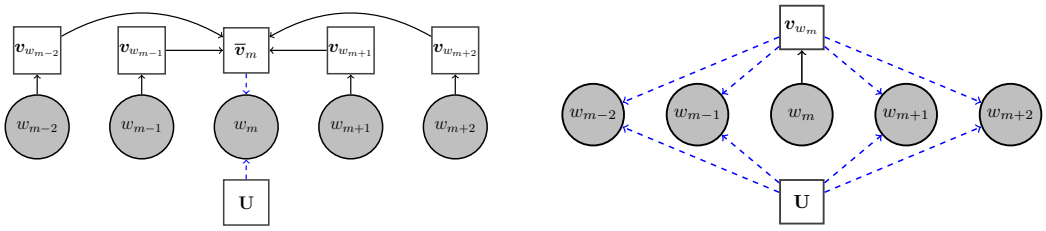
counts between clusters. After each merger, the co-occurrence vectors for the merged clusters are simply added up, so that the next optimal merger can be found efficiently.

This bottom-up procedure requires iterating over the entire vocabulary, and evaluating $K_t^2$ possible mergers at each step, where $K_t$ is the current number of clusters at step $t$ of the algorithm. Furthermore, computing the score for each merger involves a sum over $K_t^2$ clusters. The maximum number of clusters is $K_0 = V$, which occurs when every word is in its own cluster at the beginning of the algorithm. The time complexity is thus $\mathcal{O}(V^5)$.

To avoid this complexity, practical implementations use a heuristic approximation called **exchange clustering**. The $K$ most common words are placed in clusters of their own at the beginning of the process. We then consider the next most common word, and merge it with one of the existing clusters. This continues until the entire vocabulary has been incorporated, at which point the $K$ clusters are merged down to a single cluster, forming a tree. The algorithm never considers more than $K + 1$ clusters at any step, and the complexity is $\mathcal{O}(V \log V)$ in the size of the vocabulary, bounded by the cost of sorting the words at the beginning of the algorithm.

## 13.5 Neural word embeddings

Neural word embeddings combine aspects of the previous two methods: like latent semantic analysis, they are a continuous vector representation; like Brown clusters, they are trained from a likelihood-based objective. Let the vector $\boldsymbol{u}_i$ represent the $K$-dimensional




(a) Continuous bag-of-words (CBOW) (b) Skipgram

Figure 13.3: The CBOW and skipgram variants of WORD2VEC. The parameter $\mathbf{U}$ is the matrix of word embeddings, and each $\boldsymbol{v}_m$ is the context embedding for word $w_m$.

**embedding** for word $i$, and let $\boldsymbol{v}_j$ represent the $K$-dimensional embedding for context $j$. The inner product $\boldsymbol{u}_i \cdot \boldsymbol{v}_j$ represents the compatibility between word $i$ and context $j$. By incorporating this inner product into an approximation to the log-likelihood of a corpus, it is possible to estimate both parameters by backpropagation. WORD2VEC (Mikolov et al., 2013) includes two such approximations: continuous bag-of-words (CBOW) and skipgrams.

### 13.5.1 Continuous bag-of-words (CBOW)

In recurrent neural network language models, each word $w_m$ is conditioned on a recurrently-updated state vector, which is based on word representations going all the way back to the beginning of the text. The **continuous bag-of-words (CBOW)** model is a simplification: the local context is computed as an average of embeddings for words in the immediate neighborhood $m-h, m-h+1, \ldots, m+h-1, m+h$,

$$\overline{\boldsymbol{v}}_m = \frac{1}{2h} \sum_{n=1}^{h} \boldsymbol{v}_{m+n} + \boldsymbol{v}_{m-n}. \qquad [13.13]$$

Thus, CBOW is a bag-of-words model, because the order of the context words does not matter; it is continuous, because rather than conditioning on the words themselves, we condition on a continuous vector constructed from the word embeddings. The parameter $h$ determines the neighborhood size, which Mikolov et al. (2013) set to $h = 4$.

The CBOW model optimizes an approximation to the corpus log-likelihood,

$$\log \mathrm{p}(\boldsymbol{w}) \approx \sum_{m=1}^{M} \log \mathrm{p}(w_m \mid w_{m-h}, w_{m-h+1}, \ldots, w_{m+h-1}, w_{m+h}) \quad [13.14]$$

$$= \sum_{m=1}^{M} \log \frac{\exp\left(\boldsymbol{u}_{w_m} \cdot \overline{\boldsymbol{v}}_m\right)}{\sum_{j=1}^{V} \exp\left(\boldsymbol{u}_i \cdot \overline{\boldsymbol{v}}_m\right)} \quad [13.15]$$

$$= \sum_{m=1}^{M} \boldsymbol{u}_{w_m} \cdot \overline{\boldsymbol{v}}_m - \log \sum_{j=1}^{V} \exp\left(\boldsymbol{u}_i \cdot \overline{\boldsymbol{v}}_m\right). \quad [13.16]$$

### 13.5.2 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:

$$\log \mathrm{p}(\boldsymbol{w}) \approx \sum_{m=1}^{M} \sum_{n=1}^{h_m} \log \mathrm{p}(w_{m-n} \mid w_m) + \log \mathrm{p}(w_{m+n} \mid w_m) \quad [13.17]$$

$$= \sum_{m=1}^{M} \sum_{n=1}^{h_m} \log \frac{\exp(\boldsymbol{u}_{w_{m-n}} \cdot \boldsymbol{v}_{w_m})}{\sum_{j=1}^{V} \exp(\boldsymbol{u}_i \cdot \boldsymbol{v}_{w_m})} + \log \frac{\exp(\boldsymbol{u}_{w_{m+n}} \cdot \boldsymbol{v}_{w_m})}{\sum_{j=1}^{V} \exp(\boldsymbol{u}_i \cdot \boldsymbol{v}_{w_m})} \quad [13.18]$$

$$= \sum_{m=1}^{M} \sum_{n=1}^{h_m} \boldsymbol{u}_{w_{m-n}} \cdot \boldsymbol{v}_{w_m} + \boldsymbol{u}_{w_{m+n}} \cdot \boldsymbol{v}_{w_m} - 2 \log \sum_{j=1}^{V} \exp\left(\boldsymbol{u}_i \cdot \boldsymbol{v}_{w_m}\right). \quad [13.19]$$

In the skipgram approximation, each word is generated multiple times; each time it is conditioned only on a single word. This makes it possible to avoid averaging the word vectors, as in the CBOW model. The local neighborhood size $h_m$ is randomly sampled from a uniform categorical distribution over the range $\{1, 2, \ldots, h_{\text{max}}\}$; Mikolov et al. (2013) set $h_{\text{max}} = 10$. Because the neighborhood grows outward with $h$, this approach has the effect of weighting near neighbors more than distant ones. Skipgram performs better on most evaluations than CBOW (see § 13.6 for details of how to evaluate word representations), but CBOW is faster to train Mikolov et al. (2013).

### 13.5.3 Computational complexity

The WORD2VEC models can be viewed as an efficient alternative to recurrent neural network language models, which involves a recurrent state update whose time complexity is quadratic in the size of the recurrent state vector. CBOW and skipgram avoid this computation, and incur only a linear time complexity in the size of the word and context representations. However, all three models compute a normalized probability over

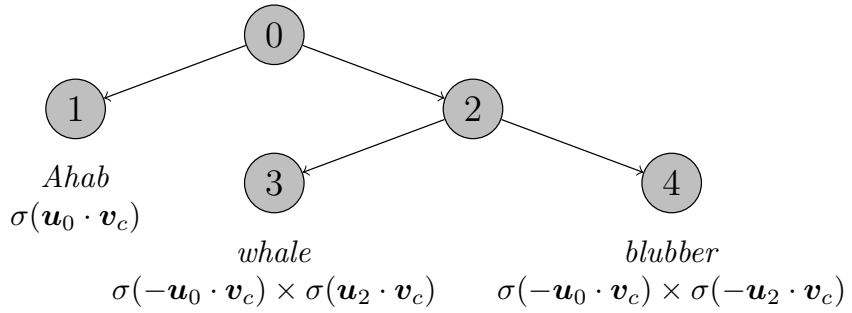



Figure 13.4: A fragment of a hierarchical softmax tree. The probability of each word is computed as a product of probabilities of local branching decisions in the tree.

word tokens; a naïve implementation of this probability requires summing over the entire 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 of these methods are also applicable to RNN language models.

**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 $\boldsymbol{v}_c$ and an output embedding associated with the node $\boldsymbol{u}_n$,

$$\Pr(\text{left at } n \mid c) = \sigma(\boldsymbol{u}_n \cdot \boldsymbol{v}_c) \quad [13.20]$$

$$\Pr(\text{right at } n \mid c) = 1 - \sigma(\boldsymbol{u}_n \cdot \boldsymbol{v}_c) = \sigma(-\boldsymbol{u}_n \cdot \boldsymbol{v}_c), \quad [13.21]$$

where $\sigma$ 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 13.4, the probability of generating each word is redefined as the product of the probabilities across its path. The sum of such probabilities is guaranteed to be one, for any context vector $\boldsymbol{v}_c \in \mathbb{R}^K$. 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 and Hinton,

2008), or by using the Huffman (1952) encoding algorithm for lossless compression (Huffman, 1952).

#### 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 difference between the score for the word that was really observed in each context, and the scores for a set of randomly selected **negative samples**:

$$\psi(i,j) = \log \sigma(\boldsymbol{u}_i \cdot \boldsymbol{v}_j) + \sum_{i' \in \mathcal{W}_{\text{neg}}} \log(1 - \sigma(\boldsymbol{u}_{i'} \cdot \boldsymbol{v}_j)), \quad [13.22]$$

where $\psi(i,j)$ is the score for word $i$ in context $j$, $\sigma$ refers to the sigmoid function, and $\mathcal{W}_{\text{neg}}$ is the set of negative samples.The objective is to maximize the sum over the corpus, $\sum_{m=1}^{M} \psi(w_m, c_m)$, where $w_m$ is token $m$ and $c_m$ is the associated context.

The set of negative samples $\mathcal{W}_{\text{neg}}$ is obtained by sampling from a unigram language model. Mikolov et al. (2013) construct this unigram language model by exponentiating the empirical word probabilities, setting $\hat{p}(i) \propto (\text{count}(i))^{\frac{3}{4}}$. This has the effect of redistributing probability mass from common to rare words. The number of negative samples increases the time complexity of training by a constant factor. Mikolov et al. (2013) report that 5-20 negative samples works for small training sets, and that two to five samples suffice for larger corpora.

### 13.5.4 Word embeddings as matrix factorization

The negative sampling objective in Equation 13.22 can be justified as an efficient approximation to the log-likelihood, but it is also closely linked to the matrix factorization objective employed in latent semantic analysis. For a matrix of word-context pairs in which all counts are non-zero, negative sampling is equivalent to factorization of the matrix $\mathbf{M}$, where $M_{ij} = \text{PMI}(i,j) - \log k$: each cell in the matrix is equal to the pointwise mutual information of the word and context, shifted by $\log k$, with $k$ equal to the number of negative samples (Levy and Goldberg, 2014). For word-context pairs that are not observed in the data, the pointwise mutual information is $-\infty$, but this can be addressed by considering only PMI values that are greater than $\log k$, resulting in a matrix of **shifted positive pointwise mutual information**,

$$M_{ij} = \max(0, \text{PMI}(i,j) - \log k). \quad [13.23]$$

Word embeddings are obtained by factoring this matrix with truncated singular value decomposition.

| word 1 | word 2 | similarity |
|---|---|---|
| *love* | *sex* | 6.77 |
| *stock* | *jaguar* | 0.92 |
| *money* | *cash* | 9.15 |
| *development* | *issue* | 3.97 |
| *lad* | *brother* | 4.46 |

Table 13.4: 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, $M_{ij} = \log \text{count}(i, j)$. We then minimize the sum of squares,

$$\begin{aligned} \min_{\boldsymbol{u},\boldsymbol{v},b,\tilde{b}} \quad & \sum_{j=1}^{V} \sum_{j \in \mathcal{C}} f(M_{ij}) \left( \widehat{\log M_{ij}} - \log M_{ij} \right)^2 \\ s.t. \quad & \widehat{\log M_{ij}} = \boldsymbol{u}_i \cdot \boldsymbol{v}_j + b_i + \tilde{b}_j, \end{aligned} \tag{13.24}$$

where $b_i$ and $\tilde{b}_j$ are offsets for word $i$ and context $j$, which are estimated jointly with the embeddings $\boldsymbol{u}$ and $\boldsymbol{v}$. The weighting function $f(M_{ij})$ is set to be zero at $M_{ij} = 0$, thus avoiding the problem of taking the logarithm of zero counts; it saturates at $M_{ij} = m_{\text{max}}$, thus avoiding the problem of overcounting common word-context pairs. This heuristic turns out to be critical to the method's performance.

The time complexity of sparse matrix reconstruction is determined by the number of non-zero word-context counts. Pennington et al. (2014) show that this number grows sublinearly with the size of the dataset: roughly $\mathcal{O}(N^{0.8})$ for typical English corpora. In contrast, the time complexity of WORD2VEC is linear in the corpus size. Computing the co-occurrence counts also requires linear time in the size of the corpus, but this operation can easily be parallelized using MapReduce-style algorithms (Dean and Ghemawat, 2008).

## 13.6 Evaluating word embeddings

Distributed word representations can be evaluated in two main ways. **Intrinsic** evaluations test whether the representations cohere with our intuitions about word meaning. **Extrinsic** evaluations test whether they are useful for downstream tasks, such as sequence labeling.

### 13.6.1 Intrinsic evaluations

A basic question for word embeddings is whether the similarity of words $i$ and $j$ is reflected in the similarity of the vectors $\boldsymbol{u}_i$ and $\boldsymbol{u}_j$. **Cosine similarity** is typically used to compare two word embeddings,

$$\cos(\boldsymbol{u}_i, \boldsymbol{u}_j) = \frac{\boldsymbol{u}_i \cdot \boldsymbol{u}_j}{||\boldsymbol{u}_i|| \times ||\boldsymbol{u}_j||}. \quad [13.25]$$

For any embedding method, we can then ask whether the cosine similarity of word embeddings is correlated with human judgments of word similarity. The WS-353 dataset (Finkelstein et al., 2002) includes similarity scores for 353 word pairs (Table 13.4). To test the accuracy of embeddings for rare and morphologically complex words, Luong et al. (2013) introduce a dataset of "rare words." Outside of English, word similarity resources are limited, mainly consisting of translations of WS-353.

Word analogies (e.g., *king:queen :: man:woman*) have also been used to evaluate word embeddings (Mikolov et al., 2013). In this evaluation, the system is provided with the first three parts of the analogy $(i_1 : j_1 :: i_2 :?)$, and the final element is predicted by finding the word embedding most similar to $\boldsymbol{u}_{i_1} - \boldsymbol{u}_{j_1} + \boldsymbol{u}_{i_2}$. Another evaluation tests whether word embeddings are related to broad lexical semantic categories called **supersenses** (Ciaramita and Johnson, 2003): verbs of motion, nouns that describe animals, nouns that describe body parts, and so on. These supersenses are annotated for English synsets in WordNet (Fellbaum, 2010). This evaluation is implemented in the `qvec` metric, which tests whether the matrix of supersenses can be reconstructed from the matrix of word embeddings (Tsvetkov et al., 2015).

Levy et al. (2015) compared several dense word representations for English — including latent semantic analysis, WORD2VEC, and GloVe — using six word similarity metrics and two analogy tasks. None of the embeddings outperformed the others on every task, but skipgrams were the most broadly competitive. Hyperparameter tuning played a key role: any method will perform badly if the wrong hyperparameters are used. Relevant hyperparameters include the embedding size, as well as algorithm-specific details such as the neighborhood size and the number of negative samples.

### 13.6.2 Extrinsic evaluations

Word representations contribute to downstream tasks like sequence labeling and document classification by enabling generalization across words. The use of distributed representations as features can be seen as a form of **semi-supervised learning**, in which performance on a supervised learning problem is augmented by learning distributed representations from unlabeled data (Miller et al., 2004; Koo et al., 2008; Turian et al., 2010). These **pre-trained 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 (§ 6.5.4). 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 scare, 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).

## 13.7 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,[4] 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 small number of times, making their embeddings unreliable. Many languages have more morphological structure than English, 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.

### 13.7.1 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:[5]

***millicuries*** This word has **morphological** structure (see § 8.1.2 for more on morphology): the prefix *milli-* indicates an amount, and the suffix *-s* indicates a plural. (A *millicurie* is an unit of radioactivity.)

[4] http://commoncrawl.org/

[5] https://code.google.com/archive/p/word2vec/, accessed September 20, 2017

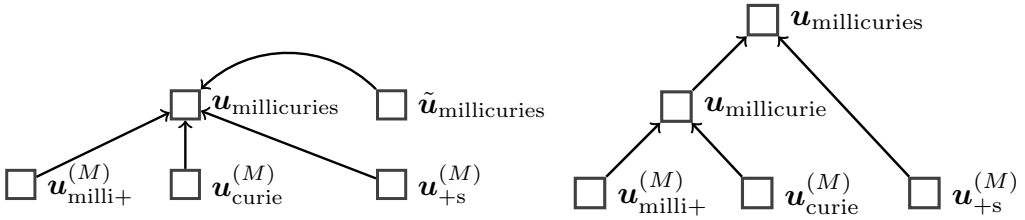



Figure 13.5: Two architectures for building word embeddings from subword units. On the left, morpheme embeddings $\boldsymbol{u}^{(m)}$ are combined by addition with the non-compositional word embedding $\tilde{\boldsymbol{u}}$ (Botha and Blunsom, 2014). On the right, morpheme embeddings are combined in a recursive neural network (Luong et al., 2013).

***caesium*** This word is a single morpheme, but the characters *-ium* are often associated with chemical elements. (*Caesium* is the British spelling of a chemical element, spelled *cesium* in American English.)

***IAEA*** This term is an acronym, as suggested by the use of capitalization. The prefix *I-* frequently refers to international organizations, and the suffix *-A* often refers to agencies or associations. (*IAEA* is the International Atomic Energy Agency.)

***Zhezhgan*** This term is in title case, suggesting the name of a person or place, and the character bigram *zh* indicates that it is likely a transliteration. (*Zhezhgan* is a mining facility in Kazakhstan.)

How can word-internal structure be incorporated into word representations? One approach is to construct word representations from embeddings of the characters or morphemes. For example, if word $i$ has morphological segments $\mathcal{M}_i$, then its embedding can be constructed by addition (Botha and Blunsom, 2014),

$$\boldsymbol{u}_i = \tilde{\boldsymbol{u}}_i + \sum_{j \in \mathcal{M}_i} \boldsymbol{u}^{(M)}_j, \tag{13.26}$$

where $\boldsymbol{u}^{(M)}_m$ is a morpheme embedding and $\tilde{\boldsymbol{u}}_i$ is a non-compositional embedding of the whole word, which is an additional free parameter of the model (Figure 13.5, left side). All embeddings are estimated from a **log-bilinear language model** (Mnih and Hinton, 2007), which is similar to the CBOW model (§ 13.5), but includes only contextual information from preceding words. The morphological segments are obtained using an unsupervised segmenter (Creutz and Lagus, 2002). For words that do not appear in the training data, the embedding can be constructed directly from the morphemes, assuming that each morpheme appears in some other word in the training data. The free parameter $\tilde{\boldsymbol{u}}$ adds flexibility: words with similar morphemes are encouraged to have similar embeddings, but this parameter makes it possible for them to be different.

Subword information can be incorporated in various ways:

**Subword units** Examples like *IAEA* and *Zhezhgan* are not based on morphological composition, and a morphological segmenter is unlikely to identify meaningful subword units for these terms. Rather than using morphemes for subword embeddings, one can use characters (Santos and Zadrozny, 2014; Ling et al., 2015; Kim et al., 2016), character $n$-grams (Wieting et al., 2016; Bojanowski et al., 2017), and **byte-pair encodings**, a compression technique which captures frequent substrings (Gage, 1994; Sennrich et al., 2016).

**Composition** Combining the subword embeddings by addition does not differentiate between orderings, nor does it identify any particular morpheme as the **root**. A range of more flexible compositional models have been considered, including recurrence (Ling et al., 2015), convolution (Santos and Zadrozny, 2014; Kim et al., 2016), and **recursive neural networks** (Luong et al., 2013), in which representations of progressively larger units are constructed over a morphological parse, e.g. ((*milli*+*curie*)+*s*), ((*in*+*flam*)+*able*), (*in*+(*vis*+*ible*)). A recursive embedding models is shown in the right panel of Figure 13.5.

**Estimation** Estimating subword embeddings from a full dataset is computationally expensive. An alternative approach is to train a subword model to match pre-trained word embeddings (Cotterell et al., 2016; Pinter et al., 2017). To train such a model, it is only necessary to iterate over the vocabulary, and the not the corpus.

### 13.7.2 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, such as synonymy (Faruqui et al., 2015). This is done by minimizing the following objective,

$$\min_{\mathbf{U}} \quad \sum_{j=1}^{V} ||\boldsymbol{u}_i - \hat{\boldsymbol{u}}_i||_2 + \sum_{(i,j)\in\mathcal{L}} \beta_{ij} ||\boldsymbol{u}_i - \boldsymbol{u}_j||_2, \tag{13.27}$$

where $\hat{\boldsymbol{u}}_i$ is the pretrained embedding of word $i$, and $\mathcal{L} = \{(i,j)\}$ is a lexicon of word relations. The hyperparameter $\beta_{ij}$ controls the importance of adjacent words having similar embeddings; Faruqui et al. (2015) set it to the inverse of the degree of word $i$, $\beta_{ij} = |\{j : (i,j) \in \mathcal{L}\}|^{-1}$. Retrofitting improves performance on a range of intrinsic evaluations, and gives small improvements on an extrinsic document classification task.

## 13.8 Distributed representations of multiword units

Can distributed semantics extended to phrases, sentences, paragraphs, and beyond? Before exploring this possibility, recall the distinction between distributed and distributional representations. Neural embeddings such as WORD2VEC are both distributed (vector-based) and distributional (derived from counts of words in context). As we consider larger units of text, the counts in any corpus must decrease: in the limit, a multi-paragraph span of text would never appear twice, except in cases of plagiarism. Thus, the meaning of a large span of text cannot be determined from distributional statistics alone; it must be computed compositionally from smaller spans. But these considerations are orthogonal to the question of whether distributed representations — dense numerical vectors — are sufficiently expressive to capture the meaning of phrases, sentences, and paragraphs.

### 13.8.1 Purely distributional methods

Some multiword phrases are non-compositional: the meaning of such phrases is not derived from the meaning of the individual words using typical compositional semantics. This includes proper nouns like *San Francisco* as well as idiomatic expressions like *kick the bucket* (Baldwin and Kim, 2010). For these cases, purely distributional approaches can work. A simple approach is to identify multiword units that appear together frequently, and then treat these units as words, learning embeddings using a technique such as WORD2VEC. The problem of identifying multiword units is sometimes called **collocation extraction**, and can be approached using metrics such as pointwise mutual information: two-word units are extracted first, and then larger units are extracted. Mikolov et al. (2013) identify such units and then treat them as words when estimating skipgram embeddings, showing that the resulting embeddings perform reasonably on a task of solving phrasal analogies, e.g. *New York : New York Times :: Baltimore : Baltimore Sun*.

### 13.8.2 Distributional-compositional hybrids

To move beyond short multiword phrases, composition is necessary. A simple but surprisingly powerful approach is to represent a sentence with the average of its word embeddings (Mitchell and Lapata, 2010). This can be considered a hybrid of the distributional and compositional approaches to semantics: the word embeddings are computed distributionally, and then the sentence representation is computed by composition. Baroni and Zamparelli (2010) embed short phrases by treating adjectives as matrices, which operate by multiplication on nouns.

The WORD2VEC approach can be stretched considerably further, embedding entire sentences using a model similar to skipgrams, in the "skip-thought" model of Kiros et al. (2015). Each sentence is **encoded** into a vector using a recurrent neural network: the encoding of sentence $t$ is set to the RNN hidden state at its final token, $\boldsymbol{h}_{M_t}^{(t)}$. This vector is

| |
|---|
| **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 13.6: 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)

then a parameter in a **decoder** model that is used to generate the previous and subsequent sentences: the decoder is another recurrent neural network, which takes the encoding of the neighboring sentence as an additional parameter in its recurrent update. (This **encoder-decoder model** is discussed at length in chapter 18.) The encoder and decoder are trained simultaneously from a likelihood-based objective, and the trained encoder can then be used to compute a distributed representation of any sentence. Skip-thought can be viewed as a hybrid of distributional and compositional approaches: the vector representation of each sentence is computed compositionally from the representations of the individual words, but the training objective is distributional, based on sentence co-occurrence across a corpus.

**Autoencoders** are a variant of encoder-decoder models in which the decoder is trained to produce the same text that was originally encoded, using only the distributed encoding vector (Li et al., 2015). By interpolating between distributed representations of two sentences, $\alpha \boldsymbol{u}_i + (1 - \alpha)\boldsymbol{u}_j$, it is possible to generate sentences that combine aspects of the two inputs, as shown in Figure 13.6 (Bowman et al., 2016). Autoencoders can also be applied to longer texts, such as paragraphs and documents. This enables applications such as **question answering**, which can be performed by matching the encoding of the question with encodings of candidate answers (Miao et al., 2016).

### 13.8.3 Supervised compositional methods

Given a supervision signal, such as a label describing the sentiment or meaning of a sentence, a wide range of compositional methods can be applied to compute a distributed representation that can then be used to predict the label. The simplest is to average the embeddings of each word in the sentence, and then pass this average through a feed-

forward neural network Iyyer et al. (2015). Convolutional and recurrent neural networks can effectively capture multiword phenomena such as negation (Kalchbrenner et al., 2014; Kim, 2014; Li et al., 2015; Tang et al., 2015). Another approach is to incorporate the syntactic structure of the sentence. **Recursive neural networks** construct a representation for each syntactic constituent by operating on the representations of the children, thus propagating the distributed representation up the tree (Socher et al., 2012). However, Li et al. (2015) show that in many cases, recurrent neural networks perform as well or better than recursive networks.

Whether convolutional, recurrent, or recursive, a key question is whether supervised sentence representations are task-specific, or whether a single supervised sentence representation model can yield useful performance on other tasks. Wieting et al. (2015) train a variety of sentence embedding models for the task of labeling pairs of sentences as **paraphrases**. They show that the resulting sentence embeddings give good performance for sentiment analysis. The **Stanford Natural Language Inference corpus** classifies sentence pairs as **entailments** (the truth of sentence $i$ implies the truth of sentence $j$), **contradictions** (the truth of sentence $i$ implies the falsity of sentence $j$), and neutral ($i$ neither entails nor contradicts $j$). Sentence embeddings trained on this dataset transfer to a wide range of classification tasks (Conneau et al., 2017).

### 13.8.4 Hybrid distributed-symbolic representations

The power of distributed representations is in their generality: the distributed representation of a unit of text can serve as a summary of its meaning, and therefore as the input for downstream tasks such as classification, matching, and retrieval. For example, distributed sentence representations can be used to recognize the paraphrase relationship between closely related sentences like the following:

(13.1) Donald thanked Vlad profusely.

(13.2) Donald conveyed to Vlad his profound appreciation.

(13.3) Vlad was showered with gratitude by Donald.

Symbolic representations are relatively brittle to this sort of variation, but are better suited to describe individual entities, the things that they do, and the things that are done to them. In examples (13.1)-(13.3), we not only know that somebody thanked someone else, but we can make a range of inferences about what has happened between the entities named *Donald* and *Vlad*. Because distributed representations do not treat entities symbolically, they lack the ability to reason about the roles played by entities across a sentence or larger discourse.[6] A hybrid between distributed and symbolic representations

[6] At 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!"

might give the best of both worlds: robustness to the many different ways of describing the same event, plus the expressiveness to support inferences about entities and the roles that they play.

A "top-down" hybrid approach is to begin with logical semantics (of the sort described in the previous two chapters), and then relax the requirement that a predefined lexicon associate words with each of a large set of predefined semantic predicates: instead, predicates are identified with distributional word clusters (Poon and Domingos, 2009; Lewis and Steedman, 2013). A "bottom-up" approach is to add minimal symbolic structure to existing distributed representations, such as vector representations for each entity (Ji and Eisenstein, 2015; Wiseman et al., 2016). This has been shown to improve performance on two problems that we will encounter in the following chapters: classification of **discourse relations** between adjacent sentences (chapter 15; Ji and Eisenstein, 2015), and **coreference resolution** of entity mentions (chapter 14; Wiseman et al., 2016; Ji et al., 2017). Research on hybrid semantic representations is still in an early stage, and future representations may deviate more boldly from existing symbolic and distributional approaches.

## Additional reading

Turney and Pantel (2010) survey a number of facets of vector word representations, focusing on matrix factorization methods. Schnabel et al. (2015) highlight problems with similarity-based evaluations of word embeddings, and present a novel evaluation that controls for word frequency. Baroni et al. (2014) address linguistic issues that arise in attempts to combine distributed and compositional representations.

In bilingual and multilingual distributed representations, embeddings are estimated for translation pairs or tuples, such as $\langle dog, perro, chien \rangle$. These embeddings can improve machine translation (Zou et al., 2013; Klementiev et al., 2012), transfer natural language processing models across languages (Täckström et al., 2012), and make monolingual word embeddings more accurate (Faruqui and Dyer, 2014). A typical approach is to learn a projection that maximizes the correlation of the distributed representations of each element in a translation pair, which can be obtained from a bilingual dictionary. Distributed representations can also be linked to perceptual information, such as image features. Bruni et al. (2014) use textual descriptions of images to obtain visual contextual information for various words, which supplements traditional distributional context. Image features can also be inserted as contextual information in log bilinear language models (Kiros et al., 2014), making it possible to automatically generate text descriptions of images.

## Exercises

1. Prove that the sum of probabilities of paths through a hierarchical softmax tree is equal to one.

2. In skipgram word embeddings, the negative sampling objective can be written as,

$$\mathcal{L} = \sum_{i \in \mathcal{V}} \sum_{j \in \mathcal{C}} \text{count}(i, j) \psi(i, j), \quad [13.28]$$

   with $\psi(i, j)$ is defined in Equation 13.22.

   Suppose we draw the negative samples from the empirical unigram distribution $\hat{\text{p}}(i) = \text{p}_{\text{unigram}}(i)$. First, compute the expectation of $\mathcal{L}$ with respect this probability.

   Next, take the derivative of this expectation with respect to the score of a single word context pair $\sigma(\boldsymbol{u}_i \cdot \boldsymbol{v}_j)$, and solve for the pointwise mutual information $\text{PMI}(i, j)$. You should be able to show that at the optimum, the PMI is a simple function of $\sigma(\boldsymbol{u}_i \cdot \boldsymbol{v}_j)$ and the number of negative samples.

3. * In Brown clustering, prove that the cluster merge that maximizes the average mutual information (Equation 13.12) also maximizes the log-likelihood objective (Equation 13.11).

For the next two problems, download a set of pre-trained word embeddings, such as the WORD2VEC or polyglot embeddings.

4. Use cosine similarity to find the most similar words to: *dog*, *whale*, *before*, *however*, *fabricate*.

5. 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 similarity.

   - *dog:puppy :: cat: ?*
   - *speak:speaker :: sing:?*
   - *France:French :: England:?*
   - *France:wine :: England:?*

The remaining problems will require you to build a classifier and test its properties. Pick a multi-class text classification dataset, such as RCV1[7]). Divide your data into training (60%), development (20%), and test sets (20%), if no such division already exists.

[7] http://www.ai.mit.edu/projects/jmlr/papers/volume5/lewis04a/lyrl2004_rcv1v2_README.htm

6. Train a convolutional neural network, with inputs set to pre-trained word embeddings from the previous problem. Use a special, fine-tuned embedding for out-of-vocabulary words. Train until performance on the development set does not improve. You can also use the development set to tune the model architecture, such as the convolution width and depth. Report $F$-measure and accuracy, as well as training time.

7. Now modify your model from the previous problem to fine-tune the word embeddings. Report $F$-measure, accuracy, and training time.

8. Try a simpler approach, in which word embeddings in the document are averaged, and then this average is passed through a feed-forward neural network. Again, use the development data to tune the model architecture. How close is the accuracy to the convolutional networks from the previous problems?

# Chapter 14

# Reference Resolution

References are one of the most noticeable forms of linguistic ambiguity, afflicting not just automated natural language processing systems, but also fluent human readers. Warnings to avoid "ambiguous pronouns" are ubiquitous in manuals and tutorials on writing style. But referential ambiguity is not limited to pronouns, as shown in the text in Figure 14.1. Each of the underlined substrings in the passage refers to an entity that is introduced earlier in the story. These references include the pronouns *he* and *his*, but also the shortened name *Cook*, and most challengingly, **nominals** such as *the firm* and *the firm's biggest growth market*.

**Reference resolution** subsumes several subtasks. This chapter will focus on **coreference resolution**, which is the task of grouping spans of text that refer to a single underlying entity, or, in some cases, a single event: for example, the spans *Tim Cook*, *he*, and *Cook* are all **coreferent**. These individual spans are called **mentions**, because they mention an entity; the entity is sometimes called the **referent**. Each mention has a set of **antecedents**, which are preceding mentions that are coreferent; for the first mention of an entity, the antecedent set is empty. The task of **pronominal anaphora resolution** requires only identifying the antecedents of pronouns. In **entity linking**, references are resolved not to other spans of text, but to entities in a knowledge base. This task is discussed in chapter 16.

Coreference resolution is a challenging problem for several reasons. Resolving different types of **referring expressions** requires different types of reasoning: the features and methods that are useful for resolving pronouns are different from those that are useful to resolve names and **nominals** (e.g., *the firm, government officials*). Coreference resolution involves not only linguistic reasoning, but also world knowledge and pragmatics: you may not have known that *China* was *Apple's biggest growth market*, but it is likely that you effortlessly resolved this reference while reading the passage in Figure 14.1.[1] A further

[1]This 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).

(14.1) *[[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 14.1: Running example (Yee and Jones, 2012). Coreferring entity mentions are underlined and bracketed.

challenge is that coreference resolution decisions are often entangled: each mention adds information about the entity, which affects other coreference decisions. This means that coreference resolution must be addressed as a structure prediction problem. But as we will see, there is no dynamic program that allows the space of coreference decisions to be searched efficiently.

## 14.1 Forms of referring expressions

There are three main forms of referring expressions — pronouns, names, and nominals.

### 14.1.1 Pronouns

Pronouns are a closed class of words that are used for references. A natural way to think about pronoun resolution is SMASH (Kehler, 2007):

- **S**earch for candidate antecedents;
- **M**atch against hard agreement constraints;
- **A**nd **S**elect using **H**euristics, which are "soft" constraints such as recency, syntactic prominence, and parallelism.

#### Search

In the search step, candidate antecedents are identified from the preceding text or speech.[2] Any noun phrase can be a candidate antecedent, and pronoun resolution usually requires

**Pragmatics** is the discipline of linguistics concerned with the formalization of such assumptions (Huang, 2015).

[2]Pronouns whose referents come later are known as **cataphora**, as in this example from Márquez (1970):

(14.1) 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.

parsing the text to identify all such noun phrases.[3] Filtering heuristics can help to prune the search space to noun phrases that are likely to be coreferent (Lee et al., 2013; Durrett and Klein, 2013). In nested noun phrases, mentions are generally considered to be the largest unit with a given head word: thus, *Apple Inc. Chief Executive Tim Cook* would be included as a mention, but *Tim Cook* would not, since they share the same head word, *Cook*.

**Matching constraints for pronouns**

References and their antecedents must agree on morphological features such as number, person, gender, and animacy. Consider the pronoun *he* in this passage from the running example:

(14.2) Tim Cook has jetted in for talks with officials as [he] seeks to clear up a pile of problems...

The pronoun and possible antecedents have the following features:

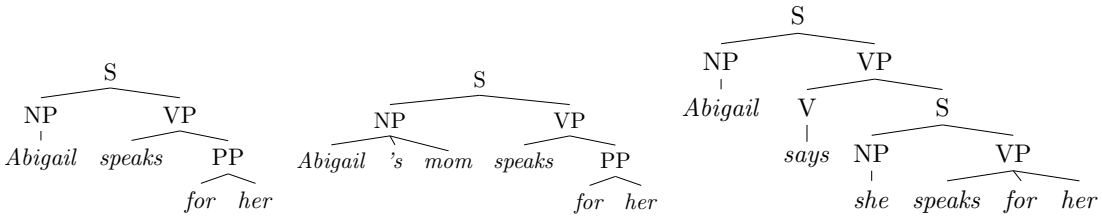
- *he*: singular, masculine, animate, third person
- *officials*: plural, animate, third person
- *talks*: plural, inanimate, third person
- *Tim Cook*: singular, masculine, animate, third person

The SMASH method searches backwards from *he*, discarding *officials* and *talks* because they do not satisfy the agreements constraints.

Another source of constraints comes from syntax — specifically, from the phrase structure trees discussed in chapter 9. Consider a parse tree in which both $x$ and $y$ are phrasal constituents. The constituent $x$ **c-commands** the constituent $y$ iff the first branching node above $x$ also dominates $y$. For example, in Figure 14.2a, *Abigail* c-commands *her*, because the first branching node above *Abigail*, S, also dominates *her*. Now, if $x$ c-commands $y$, **government and binding theory** states that $y$ can only refer to $x$ if it is a **reflexive pronoun** (e.g., *herself*). Furthermore, if $y$ is a reflexive pronoun, then its antecedent must c-command it. Thus, in Figure 14.2a, *her* cannot refer to *Abigail*; conversely, if we replace *her* with *herself*, then the reflexive pronoun must refer to *Abigail*, since this is the only candidate antecedent that c-commands it.

[3] In the OntoNotes coreference annotations, verbs can also be antecedents, if they are later referenced by nominals (Pradhan et al., 2011):

(14.1) Sales of passenger cars [grew] 22%. [The strong growth] followed year-to-year increases.




(a) (b) (c)

Figure 14.2: 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.

Now consider the example shown in Figure 14.2b. Here, *Abigail* does not c-command *her*, but *Abigail's mom* does. Thus, *her* can refer to *Abigail* — and we cannot use reflexive *herself* in this context, unless we are talking about Abigail's mom. However, *her* does not have to refer to *Abigail*. Finally, Figure 14.2c shows the how these constraints are limited. In this case, the pronoun *she* can refer to *Abigail*, because the S non-terminal puts *Abigail* outside the domain of *she*. Similarly, *her* can also refer to *Abigail*. But *she* and *her* cannot be coreferent, because *she* c-commands *her*.

**Heuristics**

After applying constraints, heuristics are applied to select among the remaining candidates. Recency is a particularly strong heuristic. All things equal, readers will prefer the more recent referent for a given pronoun, particularly when comparing referents that occur in different sentences. Jurafsky and Martin (2009) offer the following example:

(14.3) The doctor found an old map in the captain's chest. Jim found an even older map hidden on the shelf. [It] described an island.

Readers are expected to prefer the second, older map as the referent for the pronoun *it*.

However, subjects are often preferred over objects, and this can contradict the preference for recency when two candidate referents are in the same sentence. For example,

(14.4) Asha loaned Mei a book on Spanish. [She] is always trying to help people.

Here, we may prefer to link *she* to *Asha* rather than *Mei*, because of *Asha*'s position in the subject role of the preceding sentence. (Arguably, this preference would not be strong enough to select *Asha* if the second sentence were *She is visiting Argentina next month*.)

A third heuristic is parallelism:

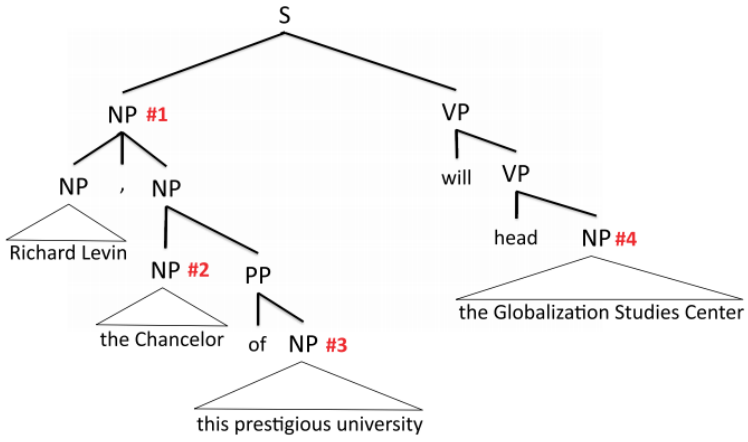



Figure 14.3: Left-to-right breadth-first tree traversal, proposed by Hobbs (1978), as implemented by Lee et al. (2013)

(14.5) Asha loaned Mei a book on Spanish. Olya loaned [her] a book on Portuguese.

Here *Mei* is preferred as the referent for *her*, contradicting the preference for the subject *Asha* in the preceding sentence.

The recency and subject role heuristics can be unified by traversing the document in a syntax-driven fashion (Hobbs, 1978): each preceding sentence is traversed breadth-first, left-to-right (Figure 14.3). This heuristic successfully handles both examples: *Asha* is preferred as the referent for *she* in (14.4), while the older map is preferred as the referent for *it* in (14.3). **Centering theory** offers an alternative unification of recency and syntactic prominence, maintaining ordered lists of candidates referents throughout the text or discourse (Grosz et al., 1995). Centering can also be viewed as a generative model, in that it predicts the form of the referring expression that will be used for each entity reference in a sentence.

In early work on reference resolution, the number of heuristics was small enough that a set of numerical weights could be set by hand (Lappin and Leass, 1994). More recent work uses machine learning to quantify the importance of each of these factors. However, pronoun resolution cannot be completely solved by morphological constraints and syntactic heuristics alone. This is shown by the classic example pair (Winograd, 1972):

(14.6) *The [city council] denied the protesters a permit because [they] feared violence.*

(14.7) *The city council denied [the protesters] a permit because [they] advocated violence.*

**Non-referential pronouns**

While pronouns are generally used for reference, they need not refer to entities. The following examples show how pronouns can refer to propositions, events, and speech acts.

(14.8) They told me that I was too ugly for show business, but I didn't believe [it].

(14.9) Asha saw Babak get angry, and I saw [it] too.

(14.10) Asha said she worked in security. I suppose [that]'s one way to put it.

These forms of reference are generally not annotated in coreference resolution datasets such as OntoNotes (Pradhan et al., 2011).

Pronouns may also have **generic referents**:

(14.11) A poor carpenter blames [her] tools.

(14.12) On the moon, [you] have to carry [your] own oxygen.

(14.13) Every farmer who owns a donkey beats [it]. (Geach, 1962)

In the OntoNotes dataset, coreference is not annotated for generic referents, even in cases like these examples, in which the same generic entity is mentioned multiple times.

Some pronouns do not refer to anything at all:

(14.14) *[It]'s raining.*
[Il] pleut. (Fr)

(14.15) [It] 's money that she's really after.

(14.16) [It] is too bad that we have to work so hard.

In the first example, *it* and *il* are **pleonastic**. The second and third examples are **cleft** and **extraposition**.[todo: explain]

How can we automatically distinguish these usages of *it* from referential pronouns? Consider the the difference between the following two examples (Bergsma et al., 2008):

(14.17) You can make [it] in advance.

(14.18) You can make [it] in showbiz.

In the second example, the pronoun *it* is non-referential. One way to see this is by substituting another pronoun, like *them*, into these examples:

(14.19) You can make [them] in advance.

(14.20) ? You can make [them] in showbiz.

The questionable grammaticality of the second example suggests that *it* is not referential. Bergsma et al. (2008) operationalize this idea by comparing distributional statistics for the $n$-grams around the word *it*, testing how often other pronouns or nouns ever appear in the same position as *it*. In cases where other pronouns are infrequent, the *it* is unlikely to be referential.

### 14.1.2 Proper Names

If a proper name is used as a referring expression, it often refers to another proper name, so that the coreference problem is simply to determine whether the two names match. Subsequent proper name references often use a shortened form, as in the running example (Figure 14.1):

(14.21) Apple Inc Chief Executive [Tim Cook] has jetted into China ... [Cook] is on his first business trip to the country ...

In this news article, the family name *Cook* is used as a referring expression; in informal conversation, it might be more typical to use the given name *Tim*, while more formal venues, such as *The Economist*, would use the title form *Mr Cook*.

A typical solution for proper name coreference is to match the syntactic **head words** of the reference with the referent. In § 9.5.2, we saw that the head word of a phrase can be identified by applying head percolation rules to the phrasal parse tree; alternatively, the head can be identified as the root of the dependency subtree covering the name. For sequences of proper nouns, the head word will be the final token.

There are a number of caveats to the practice of matching head words of proper names.

- In the European tradition, family names tend to be more specific than given names, and family names usually come last. However, other traditions have other practices: for example, in Chinese names, the family name typically comes first; in Japanese, honorifics come after the name, as in *Nobu-San* (*Mr. Nobu*).
- In organization names, the head word is often not the most informative: for example, *Georgia Tech* and *Virginia Tech* are distinguished by the modifiers *Virginia* and *Georgia*, and not the heads. Similarly, *Lebanon* does not refer to the same entity as *Southern Lebanon*, necessitating special rules for the specific case of geographical modifiers (Lee et al., 2011).
- Proper names can be nested, as in *[the CEO of [Microsoft]]*, resulting in head word match without coreference.

Despite these difficulties, proper names are the easiest category of references to resolve (Stoyanov et al., 2009). In machine learning systems, one solution is to include a

range of matching features, including exact match, head match, and string inclusion. In addition to matching features, competitive systems (e.g., Bengtson and Roth, 2008) include large lists, or **gazetteers**, of acronyms (e.g, *the National Basketball Association/NBA*), demonyms (e.g., *the Israelis/Israel*), and other aliases (e.g., *the Georgia Institute of Technology/Georgia Tech*).

### 14.1.3 Nominals

In coreference resolution, noun phrases that are neither pronouns nor names are referred to as **nominals**. In the running example (Figure 14.1), nominal references include:

- *the firm* (*Apple Inc*)
- *the firm's biggest growth market* (*China*)
- *the country* (*China*).

Nominals are generally more difficult to resolve than pronouns and names (Denis and Baldridge, 2007; Durrett and Klein, 2013), and the examples above suggest why this may be the case: world knowledge is required to identify *Apple Inc* as a *firm*, and *China* as a *growth market*. Other difficult examples include the use of colloquial expressions, such as coreference between *Clinton campaign officials* and *the Clinton camp* (Soon et al., 2001).

## 14.2 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 14.1, the ground truth coreference annotation is:

$$\boldsymbol{c}_1 = \{\textit{Apple Inc}_{1:2}, \textit{the firm}_{27:28}\} \quad [14.1]$$
$$\boldsymbol{c}_2 = \{\textit{Apple Inc Chief Executive Tim Cook}_{1:6}, \textit{he}_{17}, \textit{Cook}_{33}, \textit{his}_{36}\} \quad [14.2]$$
$$\boldsymbol{c}_3 = \{\textit{China}_{10}, \textit{the firm 's biggest growth market}_{27:32}, \textit{the country}_{40:41}\} \quad [14.3]$$

Each row specifies the token spans that mention an entity. ("Singleton" entities, which are mentioned only once (e.g., *talks*, *government officials*), are excluded.) Equivalently, if given a set of $M$ mentions, $\{m_i\}_{i=1}^{M}$, each mention $i$ can be assigned to a cluster $z_i$, where $z_i = z_j$ if $i$ and $j$ are coreferent. The cluster assignments $z$ are invariant under permutation. The unique clustering associated with the assignment $\boldsymbol{z}$ is written $\boldsymbol{c}(\boldsymbol{z})$.

[4] In 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 such as pleonastic *it*, 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.

**Mention identification** The task of identifying mention spans for coreference resolution is often performed by applying a set of heuristics to the phrase structure parse of each sentence. A typical approach is to start with all noun phrases and named entities, and then apply filtering rules to remove nested noun phrases with the same head (e.g., *[Apple CEO [Tim Cook]]*), numeric entities (e.g., *[100 miles]*, *[97%]*), pleonastic *it*, etc (Lee et al., 2013; Durrett and Klein, 2013). In general, these deterministic approaches err n favor of recall, since the mention clustering component can choose to ignore false positive mentions, but cannot recover from false negatives. An alternative is to consider all spans (up to some finite length) as candidate mentions, performing mention identification and clustering jointly (Daumé III and Marcu, 2005; Lee et al., 2017).

**Mention clustering** The overwhelming majority of research on coreference resolution addresses the subtask of mention clustering, and this will be the focus of the remainder of this chapter. There are two main sets of approaches, and as usual, they are distinguished by an independence assumption. In *mention-based models*, the scoring function for a coreference clustering decomposes over pairs of mentions. These pairwise decisions are then aggregated, using a clustering heuristic. Mention-based coreference clustering can be treated as a fairly direct application of supervised classification or ranking. However, the mention-pair independence assumption can result in incoherent clusters, like {*Hillary Clinton* ← *Clinton* ← *Mr Clinton*}, in which the pairwise links score well, but the overall result is unsatisfactory. *Entity-based models* address this issue by scoring entities holistically. This can make inference more difficult, since the number of possible entity groupings is exponential in the number of mentions.

### 14.2.1 Mention-pair models

In the **mention-pair model**, a binary label $y_{i,j} \in \{0, 1\}$ is assigned to each pair of mentions $\langle i, j \rangle$, where $i < j$. If $i$ and $j$ corefer ($z_i = z_j$), then $y_{i,j} = 1$; otherwise, $y_{i,j} = 0$. The mention *he* in Figure 14.1 is preceded by five other mentions: (1) *Apple Inc*; (2) *Apple Inc Chief Executive Tim Cook*; (3) *China*; (4) *talks*; (5) *government officials*. The correct mention 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, such as mention 3 in the example, then $y_{i,j} = 0$ for all $i$. The same is true for "mentions" that do not refer to any entity, such as pleonastic pronouns. If mention $j$ refers to an entity that has been mentioned more than once, then $y_{i,j} = 1$ for all $i < j$ that mention the referent.

By transforming coreference into a set of binary labeling problems, the mention-pair model makes it possible to apply an off-the-shelf binary classifier (Soon et al., 2001). This classifier is applied to each mention $j$ independently, searching backwards from $j$ until finding an antecedent $i$ which corefers with $j$ with high confidence. After identifying a single **antecedent**, the remaining mention pair labels can be computed by transitivity: if $y_{i,j} = 1$ and $y_{j,k} = 1$, then $y_{i,k} = 1$.

Since the ground truth annotations give entity chains $\boldsymbol{c}$ but not individual mention-pair labels $\boldsymbol{y}$, an additional heuristic must be employed to convert the labeled data into training examples for classification. A typical approach is to generate at most one positive labeled instance $y_{a_i,i} = 1$ for mention $i$, where $a_i = \max\{j : j < i \wedge z_i = z_j\}$ is the index of the most recent mention that refers to the same entity as $i$. We then generate negative labeled instances $y_{i,j} = 0$ for all $i > a_j$. In the running example, the most recent antecedent of the pronoun *he* is $a_6 = 2$, so the training data would be:

$$y_{2,6} = 1, \qquad y_{3,6} = y_{4,6} = y_{5,6} = 0. \tag{14.4}$$

The variable $y_{1,6}$ is not part of the training data, because the first mention appears before the true antecedent $a_6 = 2$.

### 14.2.2 Mention-ranking models

In **mention ranking** (Denis and Baldridge, 2007), the classifier learns to identify a single antecedent $a_i \in \{\epsilon, 1, 2, \ldots, i-1\}$ for each referring expression $i$,

$$\hat{a}_i = \underset{a \in \{\epsilon,1,2,\ldots,i-1\}}{\text{argmax}} \; \psi_M(a, i), \tag{14.5}$$

where $\psi_M(a, i)$ is a score for the mention pair $\langle a, i \rangle$. If $a = \epsilon$, then mention $i$ does not refer to any previously-introduced entity — it is not **anaphoric**. Mention-ranking is similar to the mention-pair model, but all candidates are considered simultaneously, and at most a single antecedent is selected. The mention-ranking model explicitly accounts for the possibility that mention $i$ is not anaphoric, through the score $\psi_M(\epsilon, i)$. The determination of anaphoricity can be made by a special classifier in a preprocessing step, so that non-$\epsilon$ antecedents are identified only for spans that are determined to be anaphoric (Denis and Baldridge, 2008).

As a learning problem, ranking can be trained using the same objectives as in discriminative classification. For each mention $i$, we can define a gold antecedent $a_i^*$, and an associated loss, such as the hinge loss, $\ell_i = (1 - \psi_M(a_i^*, i) + \psi_M(\hat{a}, i))_+$ or the negative log-likelihood, $\ell_i = -\log \text{p}(a_i^* \mid i; \boldsymbol{\theta})$. (For more on learning to rank, see § 16.1.1.) But as with the mention-pair model, there is a mismatch between the labeled data, which comes in the form of mention sets, and the desired supervision, which would indicate the specific antecedent of each mention. The antecedent variables $\{a_i\}_{i=1}^M$ relate to the mention sets in a many-to-one mapping: each set of antecedents induces a single clustering, but a 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. The antecedent can be treated as a latent variable, in the manner of the **latent variable**

**perceptron** from § 11.4.2 (Fernandes et al., 2014):

$$\hat{\boldsymbol{a}} = \operatorname*{argmax}_{\boldsymbol{a}} \sum_{i=1}^{M} \psi_M(a_i, i) \quad [14.6]$$

$$\boldsymbol{a}^* = \operatorname*{argmax}_{\boldsymbol{a} \in \mathcal{A}(\boldsymbol{c})} \sum_{i=1}^{M} \psi_M(a_i, i) \quad [14.7]$$

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \sum_{i=1}^{M} \boldsymbol{f}(a_i^*, i) - \boldsymbol{f}(\hat{a}_i, i), \quad [14.8]$$

where $\mathcal{A}(\boldsymbol{c})$ is the set of antecedent structures that is compatible with the ground truth coreference clustering $\boldsymbol{c}$. 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 $\boldsymbol{m}$, we compute the following probabilities:

$$\mathrm{p}(\boldsymbol{c} \mid \boldsymbol{m}) = \sum_{\boldsymbol{a} \in \mathcal{A}(\boldsymbol{c})} \mathrm{p}(\boldsymbol{a} \mid \boldsymbol{m}) = \sum_{\boldsymbol{a} \in \mathcal{A}(\boldsymbol{c})} \prod_{i=1}^{M} \mathrm{p}(a_i \mid i, \boldsymbol{m}) \quad [14.9]$$

$$\mathrm{p}(a_i \mid i, \boldsymbol{m}) = \frac{\exp\left(\psi_M(a_i, i)\right)}{\sum_{j \in \{\epsilon, 1, 2, \ldots, i-1\}} \exp\left(\psi_M(a', i)\right)}. \quad [14.10]$$

This objective rewards models that assign high scores for all valid antecedent structures. In the running example, this would correspond to summing the probabilities of the two valid antecedents for *Cook*, *he* and *Apple Inc Chief Executive Tim Cook*. In one of the exercises, you will compute the number of valid antecedent structures for a given clustering.

### 14.2.3 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:

$$m_1 = \textit{Hillary Clinton} \quad [14.11]$$

$$m_2 = \textit{Clinton} \quad [14.12]$$

$$m_3 = \textit{Bill Clinton} \quad [14.13]$$

A mention-pair system might predict $\hat{y}_{1,2} = 1, \hat{y}_{2,3} = 1, \hat{y}_{1,3} = 0$. Similarly, a mention-ranking system might choose $\hat{a}_2 = 1$ and $\hat{a}_3 = 2$. Logically, if mentions 1 and 3 are both coreferent with mention 2, then all three mentions must refer to the same entity. This constraint is known as **transitive closure**.

Transitive closure can be applied *post hoc*, revising the independent mention-pair or mention-ranking decisions. However, there are many possible ways to enforce transitive 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 documents with many mentions, there may be many violations of transitive closure, and many possible fixes. Transitive closure can be enforced by always adding edges, so that $\hat{y}_{1,3} = 1$ is preferred (e.g., Soon et al., 2001), but this can result in overclustering, with too many mentions grouped into too few entities.

Mention-pair coreference resolution can be viewed as a constrained optimization problem,

$$\max_{\boldsymbol{y} \in \{0,1\}^M} \quad \sum_{j=1}^{M} \sum_{i=1}^{j} \psi_M(i,j) \times y_{i,j}$$
$$\text{s.t.} \quad y_{i,j} + y_{j,k} - 1 \leq y_{i,k}, \quad \forall i < j < k,$$

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 § 12.2.2).

### 14.2.4 Entity-based models

A key weakness of mention-based models is that they treat coreference resolution as a classification or ranking problem, when in fact it is 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 defining 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,

$$\max_{\boldsymbol{z}} \quad \sum_{e=1} \psi_E(\{i : z_i = e\}), \tag{14.14}$$

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 4: the goal is to obtain clusters of mentions that are internally coherent. The number of possible clusterings is the **Bell number**, which is

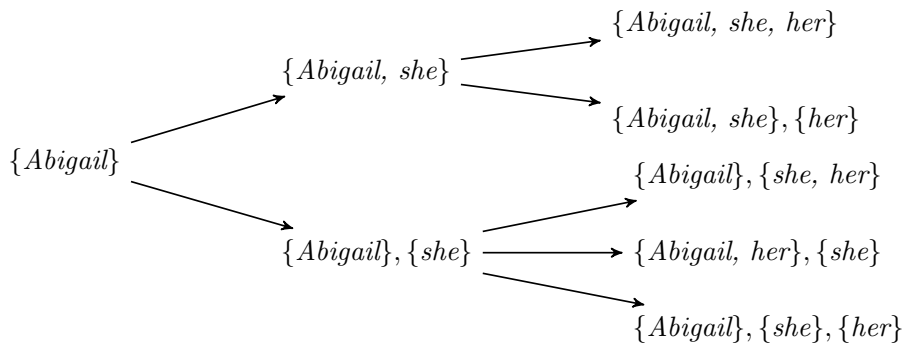



Figure 14.4: The Bell Tree for the sentence *Abigail says she cooks for her*. Which paths are excluded by the syntactic constraints mentioned in § 14.1.1?

defined by the following recurrence (Bell, 1934; Luo et al., 2004),

$$B_n = \sum_{k=0}^{n-1} B_k \binom{n-1}{k} = \frac{1}{e} \sum_{k=0}^{\infty} \frac{k^n}{k!}. \quad [14.15]$$

This recurrence is illustrated by the Bell tree, which is applied to a short coreference problem in Figure 14.4. 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 14.14. This approach is sometimes called **cluster ranking**, in contrast to mention ranking.

***Generative models of coreference** Contemporary state-of-the-art entity-based methods employ incremental clustering, but another line of research focuses on probabilistic **generative models**, in which the mentions in the document are conditioned on a set of latent entities (Haghighi and Klein, 2007, 2010). An advantage of these methods is that they can be learned from unlabeled data (Poon and Domingos, 2008, e.g.,); a disadvantage is that probabilistic inference is required not just for learning, but also for prediction. Furthermore, generative models require independence assumptions that are difficult to apply in coreference resolution, where the diverse and heterogeneous features do not admit an easy decomposition into mutually independent subsets.

#### Incremental cluster ranking

The SMASH method (§ 14.1.1) can be extended to entity-based coreference resolution by building up coreference clusters while moving through the document (Cardie and Wagstaff,

1999). At each mention, the algorithm iterates backwards through possible antecedent clusters; but unlike SMASH, a cluster is selected only if *all* members of its cluster are compatible with the current mention. As mentions are added to a cluster, so are their features (e.g., gender, number, animacy). In this way, incoherent chains like {*Hillary Clinton*, *Clinton*, *Bill Clinton*} can be avoided. However, an incorrect assignment early in the document — known as a **search error** — might make it impossible to correctly resolve references later on.

More sophisticated search strategies can help to ameliorate the risk of search errors. One approach is **beam search** (§ 10.3), in which a set of hypotheses is maintained throughout search. Each hypothesis represents a path through the Bell tree (Figure 14.4). Hypotheses are "expanded" either by adding the next mention to an existing cluster, or by starting a new cluster. Each expansion receives a score, based on Equation 14.14, and the top $K$ hypotheses are kept on the beam as the algorithm moves to the next step.

Incremental cluster ranking can be made more accurate by performing multiple passes over the document, applying rules (or "sieves") with increasing recall and decreasing precision at each pass (Lee et al., 2013). In the early passes, coreference links are proposed only between mentions that are highly likely to corefer (e.g., exact string match for full names and nominals). Information can then be shared among these mentions, so that when more permissive matching rules are applied later, agreement is preserved across the entire cluster. For example, in the case of {*Hillary Clinton*, *Clinton*, *she*}, the name-matching sieve would link *Clinton* and *Hillary Clinton*, and the pronoun-matching sieve would then link *she* to the combined cluster. A deterministic multi-pass system won nearly every track of the 2011 CoNLL shared task on coreference resolution (Pradhan et al., 2011). Given the dominance of machine learning in virtually all other areas of natural language processing — and more than fifteen years of prior work on machine learning for coreference — this was a surprising result, even if learning-based methods have subsequently regained the upper hand (e.g., Lee et al., 2017, the state-of-the-art at the time of this writing).

#### Incremental perceptron

Incremental coreference resolution can be learned with the **incremental perceptron**, as described in § 10.3.2. 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 lienar cluster ranking model,

$$\psi_E(\{i : z_i = e\}) = \sum_{i:z_i=e} \boldsymbol{\theta} \cdot \boldsymbol{f}(i, \{j : j < i \wedge z_j = e\}), \quad [14.16]$$

where the score for each cluster is computed as the sum of scores of linking decisions and $\boldsymbol{f}(i, \varnothing)$ is a set of features for the non-anaphoric mention that initiates the cluster. Using

Figure 14.4 as an example, suppose that the ground truth is,

$$\boldsymbol{c}^* = \{\textit{Abigail, her}\}, \{\textit{she}\}, \tag{14.17}$$

but that with a beam of size one, the learner reaches the hypothesis,

$$\hat{\boldsymbol{c}} = \{\textit{Abigail, she}\}. \tag{14.18}$$

This hypothesis is incompatible with $\boldsymbol{c}^*$, so an update is needed:

$$\begin{aligned}
\boldsymbol{\theta} \leftarrow & \boldsymbol{\theta} + \boldsymbol{f}(\boldsymbol{c}^*) - \boldsymbol{f}(\hat{\boldsymbol{c}}) && [14.19]\\
= & \boldsymbol{\theta} + (\boldsymbol{f}(\textit{Abigail}, \varnothing) + \boldsymbol{f}(\textit{she}, \varnothing)) - (\boldsymbol{f}(\textit{Abigail}, \varnothing) + \boldsymbol{f}(\textit{she}, \{\textit{Abigail}\})) && [14.20]\\
= & \boldsymbol{\theta} + \boldsymbol{f}(\textit{she}, \varnothing) - \boldsymbol{f}(\textit{she}, \{\textit{Abigail}\}). && [14.21]
\end{aligned}$$

This style of incremental update can also be applied to a margin loss between the gold clustering and the top clustering on the beam. By backpropagating from this loss, it is possible to train a neural network, in which the score for each entity is a function of embeddings for the entity mentions (Wiseman et al., 2015).

**Reinforcement learning**

Reinforcement learning is a topic worthy of an entire textbook in its own right (Sutton and Barto, 1998),[5] so this section will provide only a very brief overview, in the context of coreference resolution. A stochastic **policy** assigns a probability to each possible **action**, conditional on the context. The goal is to learn a policy that achieves a high expected reward, or equivalently, a low expected cost.

In incremental cluster ranking, a complete clustering on $M$ mentions can be produced by a sequence of $M$ actions, in which the action $z_i$ either merges mention $i$ with an existing cluster or begins a new cluster. We can therefore create a stochastic policy using the cluster scores (Clark and Manning, 2016),

$$\Pr(z_i = e; \boldsymbol{\theta}) = \frac{\exp \psi_E(i \cup \{j : z_j = e\}; \boldsymbol{\theta})}{\sum_{e'} \exp \psi_E(i \cup \{j : z_j = e'\}'; \boldsymbol{\theta})}, \tag{14.22}$$

where $\psi_E(i \cup \{j : z_j = e\}; \boldsymbol{\theta})$ is the score under parameters $\boldsymbol{\theta}$ for assigning mention $i$ to cluster $e$. This score can be an arbitrary function of the mention $i$, the cluster $e$ and its (possibly empty) set of mentions; it can also include the history of actions taken thus far.

[5] A 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).

If a policy assigns probability $\text{p}(\boldsymbol{c}(\boldsymbol{z}); \boldsymbol{\theta})$ to clustering $\boldsymbol{c}(\boldsymbol{z})$, then its expected loss is,

$$L(\boldsymbol{\theta}) = \sum_{\boldsymbol{c} \in \mathcal{C}(\boldsymbol{m})} \text{p}_{\boldsymbol{\theta}}(\boldsymbol{c}) \times \ell(\boldsymbol{c}), \tag{14.23}$$

where $\mathcal{C}(\boldsymbol{m})$ is the set of possible clusterings for mentions $\boldsymbol{m}$. The loss $\ell(\boldsymbol{c})$ can be based on any arbitrary scoring function, including the complex evaluation metrics used in coreference resolution (see § 14.4). This is an advantage for reinforcement learning, which can be trained directly on the evaluation metric — unlike traditional supervised learning, which requires a loss function that is differentiable and decomposable across individual decisions.

Rather than summing over the exponentially many possible clusterings, we can approximate the expectation by sampling trajectories of actions, $\boldsymbol{z} = (z_1, z_2, \ldots, z_M)$, from 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 $\boldsymbol{z}$ corresponds to a single clustering $\boldsymbol{c}$, and so we can write the loss of an action sequence as $\ell(\boldsymbol{c}(\boldsymbol{z}))$. The **policy gradient** algorithm computes the gradient of the expected loss as an expectation over trajectories (Sutton et al., 2000),

$$\frac{\partial}{\partial \boldsymbol{\theta}} L(\boldsymbol{\theta}) = E_{\boldsymbol{z} \sim \mathcal{Z}(\boldsymbol{m})} \ell(\boldsymbol{c}(\boldsymbol{z})) \sum_{i=1}^{M} \frac{\partial}{\partial \boldsymbol{\theta}} \log \text{p}(z_i \mid \boldsymbol{z}_{1:i-1}, \boldsymbol{m}) \tag{14.24}$$

$$\approx \frac{1}{K} \sum_{k=1}^{K} \ell(\boldsymbol{c}(\boldsymbol{z}^{(k)})) \sum_{i=1}^{M} \frac{\partial}{\partial \boldsymbol{\theta}} \log \text{p}(z_i^{(k)} \mid \boldsymbol{z}_{1:i-1}^{(k)}, \boldsymbol{m}) \tag{14.25}$$

$$\tag{14.26}$$

where the action sequence $\boldsymbol{z}^{(k)}$ is sampled from the current policy. Unlike the incremental perceptron, an update is not made until the complete action sequence is available.

#### Learning to search

Policy gradient can suffer from high variance: while the average loss over $K$ samples is asymptotically equal to the expected reward of a given policy, this estimate may not be accurate unless $K$ is very large. This can make it difficult to allocate credit and blame to individual actions. In **learning to search**, this problem is addressed through the addition of an **oracle** policy, which is known to receive zero or small loss. The oracle policy can be used in two ways:

- The oracle can be used to generate partial hypotheses that are likely to score well, by generating $i$ actions from the initial state. These partial hypotheses are then used as starting points for the learned policy. This is known as **roll-in**.

**Algorithm 17** Learning to search for entity-based coreference resolution

**procedure** COMPUTE-GRADIENT(mentions $\boldsymbol{m}$, loss function $\ell$, parameters $\boldsymbol{\theta}$)
  $L(\boldsymbol{\theta}) \leftarrow 0$
  $\boldsymbol{z} \sim \text{p}(\boldsymbol{z} \mid \boldsymbol{m}; \boldsymbol{\theta})$ ▷ Sample a trajectory from the current policy
  **for** $i \in \{1, 2, \ldots M\}$ **do**
    **for** action $z \in \mathcal{Z}(\boldsymbol{z}_{1:i-1}, \boldsymbol{m})$ **do** ▷ All possible actions after history $\boldsymbol{z}_{1:i-1}$
      $\boldsymbol{h} \leftarrow \boldsymbol{z}_{1:i-1} \oplus z$ ▷ Concatenate history $\boldsymbol{z}_{1:i-1}$ with action $z$
      **for** $j \in \{i+1, i+2, \ldots, M\}$ **do** ▷ Roll-out
        $h_j \leftarrow \text{argmin}_h \, \ell(\boldsymbol{h}_{1:j-1} \oplus h)$ ▷ Oracle selects action with minimum loss
      $L(\boldsymbol{\theta}) \leftarrow L(\boldsymbol{\theta}) + \text{p}(z \mid \boldsymbol{z}_{1:i-1}, \boldsymbol{m}; \boldsymbol{\theta}) \times \ell(\boldsymbol{h})$ ▷ Update expected loss
  **return** $\frac{\partial}{\partial \boldsymbol{\theta}} L(\boldsymbol{\theta})$

- The oracle can be used to compute the minimum possible loss from a given state, by generating $M - i$ actions from the current state until completion. This is known as **roll-out**.

The oracle can be combined with the existing policy during both roll-in and roll-out, sampling actions from each policy (Daumé III et al., 2009). One approach is to gradually decrease the number of actions drawn from the oracle over the course of learning (Ross et al., 2011).

In the context of entity-based coreference resolution, Clark and Manning (2016) use the learned policy for roll-in and the oracle policy for roll-out. Algorithm 17 shows how the gradients on the policy weights are computed in this case. In this application, the oracle is "noisy", because it selects the action that minimizes only the *local* loss — the accuracy of the coreference clustering up to mention $i$ — rather than identifying the action sequence that will lead to the best final coreference clustering on the entire document. When learning from noisy oracles, it can be helpful to mix in actions from the current policy with the oracle during roll-out (Chang et al., 2015).

## 14.3 Representations for coreference resolution

Historically, coreference resolution has relied on an array of hand-engineered features to capture the linguistic constraints and preferences described in § 14.1 (Soon et al., 2001). Later work has documented the utility of large feature sets, including lexical and bilexical features on mention pairs (Björkelund and Nugues, 2011; Durrett and Klein, 2013). The most recent and successful methods replace many (but not all) of these features with distributed representations of mentions and entities (Wiseman et al., 2015; Clark and Manning, 2016; Lee et al., 2017).

### 14.3.1 Features

Coreference features generally rely on a preprocessing pipeline to provide part-of-speech tagging and phrase structure parsing. This pipeline makes it possible to design features that capture many of the phenomena from § 14.1, and it is also necessary for typical approaches to mention identification. However, this pipeline may introduce errors, which can propagate to the downstream coreference clustering system. Furthermore, the existence of such a pipeline presupposes resources such as treebanks, which do not exist for many languages.[6]

#### Mention features

Features of individual mentions can help to predict anaphoricity. In systems where mention detection is performed jointly with coreference resolution, they can also predict whether a span of text is likely to be a mention. For mention $i$, typical features include:

**Mention type** Each span can be identified as a pronoun, name, or nominal, using the part-of-speech of the head word of the mention: both the Penn Treebank and Universal Dependencies tagsets (§ 7.1.1) include tags for pronouns and proper nouns, and all other heads can be marked as nominals (Haghighi and Klein, 2009).

**Mention width** The number of tokens in a mention is a rough predictor of its anaphoricity, with longer mentions being less likely to refer back to previously-defined entities.

**Lexical features** The first, last, and head words can help to predict anaphoricity; they are also useful in conjunction with features such as mention type and part-of-speech, providing a rough measure of agreement (Björkelund and Nugues, 2011). The number of lexical features can be very large, so it can be helpful to select only frequently-occurring features (Durrett and Klein, 2013).

**Morphosyntactic features** These features include the part-of-speech, number, gender, and dependency ancestors.

The features for mention $i$ and candidate antecedent $a$ can conjoined, producing joint features that can help to assess the compatibility of the two mentions. For example, Durrett and Klein (2013) conjoin each feature with the mention types of the anaphora and the

[6] The 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`

antecedent. Coreference resolution corpora such as ACE and OntoNotes contain documents from various genres. By conjoining the genre with other features, it is possible to learn genre-specific feature weights.

#### Mention-pair features

For any pair of mentions $i$ and $j$, typical features include:

**Distance** The number of intervening tokens, mentions, and sentences can all be used as distance features. These distances can be computed on the surface text, or on a transformed representation reflecting the breadth-first tree traversal (Figure 14.3). Rather than using the distances directly, they are typically binned, creating binary features.

**String match** A variety of string match features can be employed: exact match, suffix match, head match, and more complex matching rules that disregard irrelevant modifiers (Soon et al., 2001).

**Compatibility** Whether the anaphor and antecedent agree with respect to morphosyntactic attributes such as gender, number, and animacy.

**Nesting** If one mention is nested inside another (e.g., *[The President of [France]]*), they generally cannot corefer.

**Same speaker** For documents with quotations, such as news articles, personal pronouns can be resolved only by determining the speaker for each mention (Lee et al., 2013). Coreference is also more likely between mentions from the same speaker.

**Gazetteers** These features fire when the anaphor and candidate antecedent appear in a gazetteer of acronyms (e.g., *USA/United States*, *GATech/Georgia Tech*), demonyms (e.g., *Israel/Israeli*), or other aliases (e.g., *Knickerbockers/New York Knicks*).

**Lexical semantics** These features use a lexical resource such as `WordNet` to determine whether the head words of the mentions are related through synonymy, antonymy, and hypernymy (§ 3.2).

**Dependency paths** The dependency path between the anaphor and candidate antecedent can help to determine whether the pair can corefer, under the government and binding constraints described in § 14.1.1.

Comprehensive lists of mention-pair features are offered by Bengtson and Roth (2008) and Rahman and Ng (2011). Neural network approaches use far fewer mention-pair features: for example, Lee et al. (2017) include only speaker, genre, distance, and mention width features.

**Semantics** In many cases, coreference seems to require knowledge and semantic inferences, as in the running example, where we link *China* with a *country* and a *growth market*. Some of this information can be gleaned from `WordNet`, which defines a graph over **synsets** (see § 3.2). For example, one of the synsets of *China* is an instance of an `Asian_nation#1`, which in turn is a hypernym of `country#2`, a synset that includes *country*.[7] Such paths can be used to measure the similarity between concepts (Pedersen et al., 2004), and this similarity can be incorporated into coreference resolution as a feature (Ponzetto and Strube, 2006). Similar ideas can be applied to knowledge graphs induced from Wikipedia (Ponzetto and Strube, 2007). But while such approaches improve relatively simple classification-based systems, they have proven less useful when added to the current generation of techniques.[8] For example, Durrett and Klein (2013) employ a range of semantics-based features — WordNet synonymy and hypernymy relations on head words, named entity types (e.g., person, organization), and unsupervised clustering over nominal heads — but find that these features give minimal improvement over a baseline system using surface features.

#### Entity features

Features for entity-mention coreference can be generated by aggregating mention-pair features over all mentions in the candidate entity (Culotta et al., 2007; Rahman and Ng, 2011). Specifically, for each binary mention-pair feature $f(i, j)$, we compute the following entity-mention features for mention $i$ and entity $e = \{j : j < i \wedge z_j = e\}$.

- ALL-TRUE: Feature $f(i, j)$ holds for all mentions $j \in e$.
- MOST-TRUE: Feature $f(i, j)$ holds for at least half and fewer than all mentions $j \in e$.
- MOST-FALSE: Feature $f(i, j)$ holds for at least one and fewer than half of all mentions $j \in e$.
- NONE: Feature $f(i, j)$ does not hold for any mention $j \in e$.

For scalar mention-pair features, similar aggregation can be performed by computing the minimum, maximum, and median values across all mentions in the cluster. Additional entity-mention features include the number of mentions currently clustered in the entity, and ALL-X and MOST-X features for each mention type.

### 14.3.2 Distributed representations of mentions and entities

Recent work has emphasized distributed representations of both mentions and entities. One potential advantage is that pre-trained embeddings could help to capture the se-

[7] `teletype font` is used to indicate wordnet synsets, and *italics* is used to indicate strings.

[8] This 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.

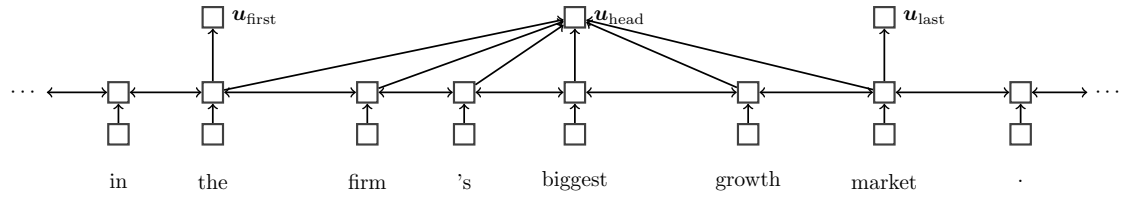



Figure 14.5: A bidirectional recurrent model of mention embeddings. The mention is represented by its first word, its last word, and an estimate of its head word, which is computed from a weighted average (Lee et al., 2017).

mantic compatibility underlying nominal coreference, helping with difficult cases like ⟨*Apple*, *the firm*⟩ and ⟨*China*, *the firm's biggest growth market*⟩. Furthermore, a distributed representation of entities can be trained to capture semantic features that are added by each mention.

### Mention embeddings

Mention embeddings are based on embeddings of the words in the mention, and the context in which those words appear (Wiseman et al., 2015). A recurrent neural network approach is shown in Figure 14.5. In this approach, a bidirectional long short-term memory (LSTM) is run over the entire document, producing hidden state representations for each word, $\{\boldsymbol{h}_m\}_{m=1}^{M}$. For each span $(s, t)$ that is a candidate mention, the distributed representation is the concatenation of four elements: a set of surface features $\boldsymbol{f}(s, t, \boldsymbol{w})$, a vector representation of the first word $\boldsymbol{u}_{\text{first}}^{(s,t)} = \boldsymbol{h}_s$, the last word $\boldsymbol{u}_{\text{last}}^{(s,t)} = \boldsymbol{h}_t$, and the "head" word $\boldsymbol{u}_{\text{head}}^{(s,t)}$. Rather than identifying the head word from the output of a phrase structure parser, the head can be computed from a neural **attention mechanism**:

$$\tilde{\alpha}_m = \boldsymbol{\theta}_\alpha \cdot \boldsymbol{h}_m \quad [14.27]$$

$$\boldsymbol{a}^{(s,t)} = \text{SoftMax}\left([\tilde{\alpha}_s, \tilde{\alpha}_{s+1}, \ldots, \tilde{\alpha}_t]\right) \quad [14.28]$$

$$\boldsymbol{u}_{\text{head}}^{(s,t)} = \sum_{m=s}^{t} a_m^{(s,t)} \boldsymbol{h}_m. \quad [14.29]$$

The vector $\boldsymbol{a}^{(s,t)}$ allocates attention across the words in the words in the span $(s, t)$; the amount of attention is constrained to sum to one by the softmax operation. This eliminates the need for syntactic parsing to recover the head word, instead learning to identify the most important word in the span (Lee et al., 2017). Attention mechanisms were introduced in neural machine translation (Bahdanau et al., 2014), and are further described in chapter 18.

Given a set of mention embeddings, each mention $i$ and candidate antecedent $a$ is scored as,

$$\psi(a, i) = \psi_S(a) + \psi_S(i) + \psi_M(a, i) \quad [14.30]$$
$$\psi_m(m) = \text{FeedForward}_S(\boldsymbol{u}_a) + \text{FeedForward}_S(\boldsymbol{u}_i) \quad [14.31]$$
$$+ \text{FeedForward}_M([\boldsymbol{u}_a; \boldsymbol{u}_i; \boldsymbol{u}_a \odot \boldsymbol{u}_i; \boldsymbol{f}(a, i, \boldsymbol{w})]), \quad [14.32]$$

where $\text{FeedForward}_S(\boldsymbol{u}_i)$ is a feedforward neural network applied to the mention representation $\boldsymbol{u}_i$. The feature vector $\boldsymbol{f}(a, i, \boldsymbol{w})$ describes mentions $a$ and $i$, including distance, speaker, and genre information. The final term is a feedforward network applied to a vector that vertically concatenates the representations of each mention, their elementwise product $\boldsymbol{u}_a \odot \boldsymbol{u}_i$, and the surface features. 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 several 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 $\boldsymbol{u}_{i,j}$ is used in the scoring function of a mention-ranking model, $\psi_M(i, j) = \boldsymbol{\theta} \cdot \boldsymbol{u}_{i,j}$. 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, $\boldsymbol{v}_e$. Each candidate mention receives an embedding $\boldsymbol{u}_i$; 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, $\psi_E(i, e) = \text{Feedforward}([\boldsymbol{v}_e; \boldsymbol{u}_i])$. If $i$ is added to entity $e$, then its representation is updated recurrently, $\boldsymbol{v}_e \leftarrow f(\boldsymbol{v}_e, \boldsymbol{u}_i)$, using a recurrent neural network such as a long short-term memory (LSTM; chapter 5). Alternatively, we can apply a **pooling** operation, such as max-pooling or average-pooling (chapter 2), setting $\boldsymbol{v}_e \leftarrow \text{Pool}(\boldsymbol{v}_e, \boldsymbol{u}_i)$. In either case, the update to the representation of entity $e$ can be thought of as adding new information about the entity from mention $i$.

## 14.4 Additional reading

**Historical notes** 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).

**Evaluating coreference resolution** The state of coreference evaluation is aggravatingly complex. Early attempts at simple evaluation metrics were found to under-penalize 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`.

## 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). That is, group the noun phrases into entities, where each entity corresponds to a set of 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:

   - The number of times new entities are introduced by each of the three types of referring expressions: pronouns, proper nouns, and nominals. Include "singleton" entities that are mentioned only once.
   - For each type of referring expression, compute the fraction of mentions that are anaphoric.

3. Apply a simple heuristic to all pronouns in the article from the previous exercise. Specifically, link each pronoun to the closest preceding noun phrase that agrees in gender, number, animacy, and person. Compute the following evaluation:

   - True positive: a pronoun that is linked to a noun phrase with which it is coreferent, or is correctly labeled as the first mention of an entity.
   - False positive: a pronoun that is linked to a noun phrase with which it is not coreferent. (This includes mistakenly linking singleton or non-referential pronouns.)
   - False negative: a pronoun that is not linked to a noun phrase with which it is coreferent.

   Compute the $F$-measure for your method, and for a trivial baseline in which every mention is its own entity. Are there any additional heuristics that would have improved the performance of this method?

4. Durrett and Klein (2013) compute the probability of the gold coreference clustering by summing over all antecedent structures that are compatible with the clustering. Compute the number of antecedent structures for a single entity with $K$ mentions.

5. Use the policy gradient algorithm to compute the gradient for the following scenario, based on the Bell tree in Figure 14.4:

   - The gold clustering $\boldsymbol{c}^*$ is $\{\textit{Abigail}, \textit{her}\}, \{\textit{she}\}$.
   - Drawing a single sequence of actions ($K = 1$) from the current policy, you obtain the following incremental clusterings:

   $$\begin{aligned}\boldsymbol{c}(a_1) &= \{\textit{Abigail}\} \\ \boldsymbol{c}(\boldsymbol{a}_{1:2}) &= \{\textit{Abigail}, \textit{she}\} \\ \boldsymbol{c}(\boldsymbol{a}_{1:3}) &= \{\textit{Abigail}, \textit{she}\}, \{\textit{her}\}.\end{aligned}$$

   - At each mention $t$, the action space $\mathcal{A}_t$ is to merge the mention with each existing cluster, or the empty cluster, with probability,

   $$\Pr(\text{Merge}(m_t, \boldsymbol{c}(\boldsymbol{a}_{1:t-1}))) \propto \exp \psi_E(m_t \cup \boldsymbol{c}(\boldsymbol{a}_{1:t-1})), \qquad [14.33]$$

   where the cluster score $\psi_E(m_t \cup c)$ is defined in Equation 14.16.

   Compute the gradient $\frac{\partial}{\partial \boldsymbol{\theta}} L(\boldsymbol{\theta})$ in terms of the loss $\ell(\boldsymbol{c}(\boldsymbol{a}))$ and the features of each (potential) cluster. Explain the differences between the gradient-based update $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \frac{\partial}{\partial \boldsymbol{\theta}} L(\boldsymbol{\theta})$ and the incremental perceptron update from this sample example.

# Chapter 15

# Discourse

## 15.1 Discourse relations in the Penn Discourse Treebank

- introduce discourse relations
- PDTB annotation framework in D-LTAG
- PDTB parsing

## 15.2 Rhetorical Structure Theory

- Higher-level discourse structure
- Shift-reduce parsing
- Applications to summarization

## 15.3 Centering

- Pronouns, forms of reference
- Smooth/rough transitions
- Entity grid implementation

# Part IV

# Applications

# Chapter 16

# Information extraction

Computers offer powerful capabilities for searching and reasoning about structured records and relational data. Some have even argued that the most important limitation of contemporary artificial intelligence is not inference or learning, but simply having too little knowledge (Lenat et al., 1990). Natural language processing provides an appealing solution: automatically construct a structured **knowledge base** by reading natural language text.

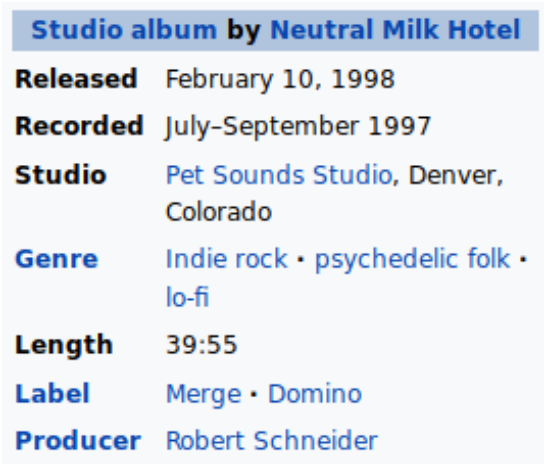
Many Wikipedia pages have an "infobox" that provides structured information about the an entity or event. An example is shown in Figure 16.1a: each row represents one or more properties of the entity IN THE AEROPLANE OVER THE SEA, a record album. The set of properties is determined by a predefined **schema**, which applies to all record albums in Wikipedia. As shown in Figure 16.1b, the values for many of these fields are indicated directly in the first few sentences of text on the same Wikipedia page.

The task of automatically constructing (or "populating") an infobox based on the text is an example of **information extraction**. Much of information extraction can be described in terms of **entities**, **relations**, and **events**.

- **Entities** are uniquely specified objects in the world, such as people (`Jeff Mangum`), places (`Athens, Georgia`), organizations (`Merge Records`), and times (`February 10, 1998`). In chapter 7, we encountered the task of **named entity recognition**, which labels tokens as parts of entity spans. In information extraction, we must go further, **linking** each entity **mention** to an element in a **knowledge base**.

- **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

(16.1) In the Aeroplane Over the Sea is the second and final studio album by the American indie rock band Neutral Milk Hotel.

(16.2) 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.

(16.3) 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 16.1: From the Wikipedia page for the album "In the Aeroplane Over the Sea", retrieved October 26, 2017.

of the album described in Figure 16.1 is described by the event,

$$\langle \textsc{title} : \texttt{In the Aeroplane Over the Sea}, \\ \textsc{artist} : \texttt{Neutral Milk Hotel}, \\ \textsc{release-date} : \texttt{1998-Feb-10}, \ldots \rangle$$

The set of arguments for an event type is defined by a **schema**. Events often refer to time-delimited occurrences: weddings, protests, purchases, terrorist attacks.

Information extraction is similar to predicate-argument semantic parsing tasks, such as semantic role labeling (chapter 12): we may think of predicates as corresponding to events, and the arguments as defining slots in the event representation. However, the goals of information extraction are usually more limited. Rather than accurately parsing every sentence, information extraction systems often focus on recognizing a few key relation or event types, or on the task of identifying all properties of a given entity. Information extraction is often evaluated by the correctness of the resulting knowledge base, and not by how many sentences were accurately parsed. Many relations and events will be mentioned multiple times in a corpus, but in information extraction, we are usually interested in identifying each relation and event only once — thus the goal here is sometimes described as **macro-reading**, as opposed to **micro-reading**, in which each sentence

must be analyzed correctly. Macro-reading systems are not penalized for ignoring difficult sentences, as long as they can recover the same information from other, easier-to-read sources. However, macro-reading systems must resolve apparent inconsistencies (was the album released on `Merge Records` or `Blue Rose Records`?), requiring reasoning across the entire dataset.

In addition to the basic tasks of recognizing entities, relations, and events, accurate information extraction systems must handle negation, and must be able to distinguish statements of fact from hopes, fears, hunches, and hypotheticals. Finally, information extraction is often paired with the problem of **question answering**, which requires accurately parsing a query, and then selecting or generating a textual answer. Question answering systems can be built on knowledge bases that are extracted from large text corpora, or may attempt to identify answers directly from the source texts.

## 16.1 Entities

The starting point for information extraction is to identify mentions of entities in text. Consider the following text:

(16.4) *The United States Army captured a hill overlooking Atlanta on May 14, 1864.*

For this sentence, we have two goals:

1. **Identify** the spans *United States Army*, *Atlanta*, and *May 14, 1864* as entity mentions. (The hill is not uniquely identified, so it is not a *named* entity.) We may also want to recognize the **named entity types**: organization, location, and date. This is **named entity recognition**, and is described in chapter 7.
2. **Link** these spans to entities in a knowledge base: `U.S. Army`, `Atlanta`, and `1864-May-14`. This task is known as **entity linking**.

The strings to be linked to entities are known as **mentions** — similar to the use of this term in coreference resolution.

- In some formulations of the entity linking task, only named entities are candidates for linking. This is sometimes called **named entity linking** (Ling et al., 2015).
- In other formulations, such as **Wikification** (Milne and Witten, 2008), any string can be a mention.

The set of target entities often corresponds to Wikipedia pages, and Wikipedia is the basis for more comprehensive knowledge bases such as YAGO (Suchanek et al., 2007), DBPedia (Auer et al., 2007), and Freebase (Bollacker et al., 2008). Entity linking may also be

performed in more "closed" settings, where a much smaller list of targets is provided in advance. The system must also determine if a mention does not refer to any entity in the knowledge base, sometimes called a **NIL entity** (McNamee and Dang, 2009).

Returning to (16.4), the three entity mentions may seem unambiguous. But the Wikipedia disambiguation page for the string *Atlanta* says otherwise:[1] there are more than twenty different towns and cities, five United States Navy vessels, a magazine, a television show, a band, and a singer — each prominent enough to have its own Wikipedia page. We now consider how to choose among these dozens of possibilities. In this chapter we will focus on supervised approaches. Unsupervised entity linking is closely related to the problem of **cross-document coreference resolution** (Bagga and Baldwin, 1998b; Singh et al., 2011).

### 16.1.1 Entity linking by learning to rank

Entity linking is often formulated as a **ranking** problem,

$$\hat{y} = \underset{y \in \mathcal{Y}(\boldsymbol{x})}{\text{argmax}}\, \psi(y, \boldsymbol{x}, \boldsymbol{c}), \tag{16.1}$$

where $y$ is a target entity, $\boldsymbol{x}$ is a description of the mention, $\mathcal{Y}(\boldsymbol{x})$ is a set of candidate entities, and $\boldsymbol{c}$ is a description of the context — such as the other text in the document, or its metadata. The function $\psi$ is a scoring function, which could be a linear model, $\psi(y, \boldsymbol{x}, \boldsymbol{c}) = \boldsymbol{\theta} \cdot \boldsymbol{f}(y, \boldsymbol{x}, \boldsymbol{c})$, or a more complex function such as a neural network. In either case, the scoring function can be learned by minimizing a margin-based **ranking loss**,

$$\ell(\hat{y}, y^{(i)}, \boldsymbol{x}^{(i)}, \boldsymbol{c}^{(i)}) = \Big(\psi(\hat{y}, \boldsymbol{x}^{(i)}, \boldsymbol{c}^{(i)}) - \psi(y^{(i)}, \boldsymbol{x}^{(i)}, \boldsymbol{c}^{(i)}) + 1\Big)_+, \tag{16.2}$$

where $y^{(i)}$ is the ground truth and $\hat{y} \neq y^{(i)}$ is the predicted target for mention $\boldsymbol{x}^{(i)}$ in context $\boldsymbol{c}^{(i)}$ (Joachims, 2002; Dredze et al., 2010).

**Candidate identification** For computational tractability, it is helpful to restrict the set of candidates, $\mathcal{Y}(\boldsymbol{x}^{(i)})$. One approach is to use a **name dictionary**, which maps from strings to the entities that they might mention. This mapping is many-to-many: a string such as *Atlanta* can refer to multiple entities, and conversely, an entity such as `Atlanta` can be referenced by multiple strings. A name dictionary can be extracted from Wikipedia, with links between each Wikipedia entity page and the anchor text of all hyperlinks that point to the page (Bunescu and Pasca, 2006; Ratinov et al., 2011). To improve recall, the name dictionary can be augmented by partial and approximate matching (Dredze et al., 2010), but as the set of candidates grows, the risk of false positives (and low precision) increases. For example, the string *Atlanta* is a partial match to *the Atlanta Fed* (a name for the `Federal Reserve Bank of Atlanta`), and a noisy match (edit distance of one) from *Atalanta* (a heroine in Greek mythology and an Italian soccer team).

[1] `https://en.wikipedia.org/wiki/Atlanta_(disambiguation)`, retrieved November 1, 2017.

**Features** Feature-based approaches to entity ranking rely on three main types of local 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 context $\boldsymbol{c}^{(i)}$, 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). Contextual similarity can be modeled using the cosine similarity of bag-of-words vectors, typically weighted using **inverse document frequency** to emphasize rare words.[2]

**Neural entity linking** An alternative approach to entity ranking 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,

$$\psi(y, \boldsymbol{x}, \boldsymbol{c}) = \boldsymbol{v}_y^\top \mathbf{W}^{(y,x)} \boldsymbol{x} + \boldsymbol{v}_y^\top \mathbf{W}^{(y,c)} \boldsymbol{c}, \quad [16.3]$$

with $\boldsymbol{v}_y \in \mathbb{R}^{K_y}$ as the vector embedding of entity $y$, $\boldsymbol{x} \in \mathbb{R}^{K_x}$ as the embedding of the mention, $\boldsymbol{c} \in \mathbb{R}^{K_c}$ as the embedding of the context, and the matrices $\mathbf{W}^{(y,x)}$ and $\mathbf{W}^{(y,c)}$ 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.

[2]The **document frequency** of word $j$ is $\text{DF}(j) = \frac{1}{N} \sum_{i=1}^{N} \delta\left(x_j^{(i)} > 0\right)$, 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** $\frac{1}{\text{DF}(j)}$ or $\log \frac{1}{\text{DF}(j)}$, to emphasize rare words (Spärck Jones, 1972).

- 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 on the text of wikipedia, with hyperlinks substituted for the anchor text.[3]
- The embedding of the mention $\boldsymbol{x}$ can be computed by averaging the embeddings of the words in the mention (Yang et al., 2016), or by one of the compositional techniques described in § 13.8.
- The embedding of the context $\boldsymbol{c}$ can be represented by a low-dimensional vector. In an auto-encoding framework, this vector is the result of noisy compression, so that it is possible to approximately reconstruct the original document (Vincent et al., 2010; Kingma and Welling, 2014). He et al. (2013) apply this idea to entity linking. The vector $\boldsymbol{c}$ can also be obtained by convolution on the embeddings of words in the document (Sun et al., 2015), or by examining metadata such as the author's social network (Yang et al., 2016).

The remaining parameters $\mathbf{W}^{(y,x)}$ and $\mathbf{W}^{(y,c)}$ can be trained by backpropagation from the margin loss in Equation 16.2.

### 16.1.2 Collective entity linking

Consider the following lists:

(16.5) California, Oregon, Washington

(16.6) Baltimore, Washington, Philadelphia

(16.7) Washington, Adams, Jefferson

In each case, the term *Washington* refers to a different entity, and this reference is strongly suggested by the other entries on the list. In the last list, all three names are ambiguous in isolation — there are dozens of other *Adams* and *Jefferson* entities in Wikipedia. But a preference for coherence motivates **collectively** linking these references to the first three U.S. presidents.

A general approach to collective entity linking is to introduce a compatibility score $\psi_c(\boldsymbol{y})$. Collective entity linking is then performed by optimizing the global objective,

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y} \in \mathbb{Y}(\boldsymbol{x})}{\text{argmax}} \, \psi_c(\boldsymbol{y}) + \sum_{i=1}^{N} \psi_\ell(y^{(i)}, \boldsymbol{x}^{(i)}, \boldsymbol{c}^{(i)}), \quad [16.4]$$

where $\mathbb{Y}(\boldsymbol{x})$ is the set of all possible collective entity assignments for the mentions in $\boldsymbol{x}$, and $\psi_\ell$ is the local scoring function for each entity $i$. For simplicity, the compatibility func-

[3] Pre-trained entity embeddings can be downloaded from `https://code.google.com/archive/p/word2vec/`.

tion is typically decomposed into a sum of pairwise scores, $\psi_c(\boldsymbol{y}) = \sum_{i=1}^{N} \sum_{j \neq i}^{N} \psi_c(y^{(i)}, y^{(j)})$. The compatibility can then be computed in a number of different ways:

- Wikipedia defines high-level categories for entities (e.g., *living people*, *Presidents of the United States*, *States of the United States*), and $\psi_c$ can reward entity pairs for the number of categories that they have in common (Cucerzan, 2007).
- Compatibility can be measured by the number of incoming hyperlinks shared by the Wikipedia pages for the two entities (Milne and Witten, 2008).
- In a neural architecture, the compatibility of two entities can be set equal to the inner product of their embeddings, $\psi_c(y^{(i)}, y^{(j)}) = \boldsymbol{v}_{y^{(i)}} \cdot \boldsymbol{v}_{y^{(j)}}$.
- A non-pairwise compatibility score can be defined using a type of latent variable model known as a **probabilistic topic model** (Blei et al., 2003; Blei, 2012). In this framework, each latent topic is a probability distribution over entities, and each document has a probability distribution over topics. Each entity helps to determine the document's distribution over topics, and in turn these topics help to resolve ambiguous entity mentions (Newman et al., 2006). Inference can be performed using the techniques described in chapter 4.

Unfortunately, collective entity linking is **NP-hard** even for pairwise compatibility functions. This means that it almost certainly cannot be solved efficiently (see Appendix B). Various approximate inference techniques have been proposed, including **integer linear programming** (Cheng and Roth, 2013), **Gibbs sampling** (Han and Sun, 2012), and graph-based algorithms (Hoffart et al., 2011; Han et al., 2011).

### 16.1.3 *Pairwise ranking loss functions

The loss function defined in Equation 16.2 considers only the highest-scoring prediction $\hat{y}$, but in fact, the true entity $y^{(i)}$ should outscore *all* other entities. A loss function based on this idea would give a gradient against the features or representations of several entities, not just the top-scoring prediction. Usunier et al. (2009) define a general ranking error function,

$$L_{\text{rank}}(k) = \sum_{j=1}^{k} \alpha_j, \quad \text{with } \alpha_1 \geq \alpha_2 \geq \cdots \geq 0, \tag{16.5}$$

where $k$ is equal to the number of labels ranked higher than the correct label $y^{(i)}$. This function defines a class of ranking errors: if $\alpha_j = 1$ for all $j$, then the ranking error is equal to the rank of the correct entity; if $\alpha_1 = 1$ and $\alpha_{j>1} = 0$, then the ranking error is one whenever the correct entity is not ranked first; if $\alpha_j$ decreases smoothly with $j$, as in $\alpha_j = \frac{1}{j}$, then the error is between these two extremes.

**Algorithm 18** WARP approximate ranking loss

**procedure** WARP($y^{(i)}, \boldsymbol{x}^{(i)}$)
  $N \leftarrow 0$
  **repeat**
    Randomly sample $y \sim \mathcal{Y}(\boldsymbol{x}^{(i)})$
    $N \leftarrow N + 1$
    **if** $\psi(y, \boldsymbol{x}^{(i)}) + 1 > \psi(y^{(i)}, \boldsymbol{x}^{(i)})$ **then** ▷ check for margin violation
      $r \leftarrow \left\lfloor |\mathcal{Y}(\boldsymbol{x}^{(i)})|/N \right\rfloor$ ▷ compute approximate rank
      **return** $L_{\text{rank}}(r) \times (\psi(y, \boldsymbol{x}^{(i)}) + 1 - \psi(y^{(i)}, \boldsymbol{x}^{(i)}))$
  **until** $N \geq |\mathcal{Y}(\boldsymbol{x}^{(i)})| - 1$ ▷ no violation found
  **return** 0 ▷ return zero loss

This ranking error will 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,

$$r(y^{(i)}, \boldsymbol{x}^{(i)}) \triangleq \sum_{y \in \mathcal{Y}(\boldsymbol{x}^{(i)}) \setminus y^{(i)}} \delta\left(1 + \psi(y, \boldsymbol{x}^{(i)}) \geq \psi(y^{(i)}, \boldsymbol{x}^{(i)})\right), \quad [16.6]$$

where $\delta(\cdot)$ is a delta function, and $\mathcal{Y}(\boldsymbol{x}^{(i)}) \setminus y^{(i)}$ is the set of all entity candidates minus the true entity $y^{(i)}$. The margin-augmented rank is the rank of the true entity, after augmenting every other candidate with a margin of one, under the current scoring function $\psi$. (The context $\boldsymbol{c}$ is omitted for clarity, and can be considered part of $\boldsymbol{x}$.)

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,

$$\ell(y^{(i)}, \boldsymbol{x}^{(i)}) = \frac{L_{\text{rank}}(r(y^{(i)}, \boldsymbol{x}^{(i)}))}{r(y^{(i)}, \boldsymbol{x}^{(i)})} \sum_{y \in \mathcal{Y}(\boldsymbol{x}) \setminus y^{(i)}} \left(\psi(y, \boldsymbol{x}^{(i)}) - \psi(y^{(i)}, \boldsymbol{x}^{(i)}) + 1\right)_+, \quad [16.7]$$

The sum in Equation 16.7 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 16.7 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, $\psi(y, \boldsymbol{x}^{(i)}) + 1 \geq \psi(y^{(i)}, \boldsymbol{x}^{(i)})$. The number of samples $N$ required to find

| | |
|---|---|
| CAUSE-EFFECT | *those cancers were caused by radiation exposures* |
| INSTRUMENT-AGENCY | *phone operator* |
| PRODUCT-PRODUCER | *a factory manufactures suits* |
| CONTENT-CONTAINER | *a bottle of honey was weighed* |
| ENTITY-ORIGIN | *letters from foreign countries* |
| ENTITY-DESTINATION | *the boy went to bed* |
| COMPONENT-WHOLE | *my apartment has a large kitchen* |
| MEMBER-COLLECTION | *there are many trees in the forest* |
| COMMUNICATION-TOPIC | *the lecture was about semantics* |

Table 16.1: Relations and example sentences from the SemEval-2010 dataset (Hendrickx et al., 2009)

such a violation yields an approximation of the margin-augmented rank of the true entity, $r(y^{(i)}, \boldsymbol{x}^{(i)}) \approx \left\lfloor \frac{|\mathcal{Y}(\boldsymbol{x})|}{N} \right\rfloor$. If a violation is found immediately, $N = 1$, the correct entity probably ranks below many others, $r \approx |\mathcal{Y}(\boldsymbol{x})|$. If many samples are required before a violation is found, $N \rightarrow |\mathcal{Y}(\boldsymbol{x})|$, 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 $\mathcal{Y}(\boldsymbol{x}^{(i)})$ is large. Note the similarity to the **negative sampling** objective in WORD2VEC (chapter 13).

## 16.2 Relations

Consider the following example:

(16.8) 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,

$$\text{PHYSICAL.LOCATED}(\texttt{George Bush}, \texttt{France}). \quad [16.8]$$

Relations take exactly two arguments, and the order of the arguments matters: the converse relation would imply that `France` is inside of `George Bush`, which is completely different.

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):

(16.9) The cup contained tea from dried ginseng.

This sentence describes a relation of type ENTITY-ORIGIN between *tea* and *ginseng*. Nominal relation extraction is closely related to **semantic role labeling** (chapter 12). The key difference is that relation extraction is restricted to a relatively small number of relation types; for example, Table 16.1 shows the ten relation types from SemEval-2010.

### 16.2.1 Pattern-based relation extraction

Early work on relation extraction focused on hand-crafted patterns (Hearst, 1992). For example, the appositive *Starbuck, a native of Nantucket* signals the relation ENTITY-ORIGIN between *Starbuck* and *Nantucket*. This pattern can be written as,

$$\text{PERSON , } \textit{a native of} \text{ LOCATION} \Rightarrow \text{ENTITY-ORIGIN}(\text{PERSON}, \text{LOCATION}). \quad [16.9]$$

This pattern will be "triggered" whenever the literal string *, a native of* occurs between an entity of type PERSON and an entity of type LOCATION. Such patterns can be generalized beyond literal matches using techniques such as lemmatization, which would enable the words (*buy*, *buys*, *buying*) to trigger the same patterns (see § 3.3.1). A more aggressive strategy would be to group all words in a WordNet synset (§ 3.2), grouping words like *buy* and *purchase*.

Relation extraction patterns can be implemented in finite-state automata (§ 8.1). If the named entity recognizer is also a finite-state machine, then the systems can be combined by finite-state transduction (Hobbs et al., 1997). This makes it possible to propagate uncertainty through the finite-state cascade. Suppose the entity recognizer hypothesizes that *Starbuck* refers to either a PERSON or a LOCATION; in the composed transducer, the relation extractor would be free to select the PERSON annotation when it appears in the context of an appropriate pattern.

### 16.2.2 Relation extraction as a classification task

Relation extraction can be formulated as a classification problem,

$$\hat{r}_{(i,j),(m,n)} = \underset{r \in \mathcal{R}}{\text{argmax}}\, \psi(r, (i,j), (m,n), \boldsymbol{w}), \quad [16.10]$$

where $r \in \mathcal{R}$ is a relation type (possibly NIL), $\boldsymbol{w}_{i:j}$ is the span of the first argument, and $\boldsymbol{w}_{m:n}$ is the span of the second argument. The argument $\boldsymbol{w}_{m:n}$ may appear before or after $\boldsymbol{w}_{i:j}$ in the text, or they may overlap; we stipulate only that $\boldsymbol{w}_{i:j}$ is the first argument of the relation. We now consider three alternatives for computing the scoring function $\psi$.

### Feature-based classification

To build a feature-based classifier, we define the scoring function as,

$$\psi(r, (i,j), (m,n), \boldsymbol{w}) = \boldsymbol{\theta} \cdot \boldsymbol{f}(r, (i,j), (m,n), \boldsymbol{w}), \quad [16.11]$$

with $\boldsymbol{\theta}$ representing a vector of weights, and $\boldsymbol{f}(\cdot)$ a vector of features. The pattern-based methods described in § 16.2.1 suggest several features:

- Local features of $\boldsymbol{w}_{i:j}$ and $\boldsymbol{w}_{m:n}$, including: the strings themselves; whether they are recognized as entities, and if so, which type; whether the strings are present in a **gazetteer** of entity names; each string's syntactic **head** (§ 8.2.2).
- Features of the span between the two arguments, $\boldsymbol{w}_{j:m}$ or $\boldsymbol{w}_{n:i}$ (depending on which argument appears first): the length of the span; the specific words that appear in the span, either as a literal sequence or a bag-of-words; the wordnet synsets (§ 3.2) that appear in the span between the arguments.
- Features of the syntactic relationship between the two arguments, typically the **dependency path** between the arguments (§ 12.2.1). Example dependency paths are shown in Table 16.2.

### Kernels

Suppose that the first line of Table 16.2 is a labeled example, and the remaining lines are instances to be classified. A feature-based approach would have to decompose the dependency paths into features that capture individual edges, with or without their labels, and then learn weights for each of these features: for example, the second line contains identical dependencies, but different arguments; the third line contains a different inflection of the word *travel*; the fourth and fifth lines each contain an additional edge on the dependency path; and the sixth example uses an entirely different path. Rather than attempting to create local features that capture all of the ways in which these dependencies paths are similar and different, we can instead define a similarity function $\kappa$, which computes a score for any pair of instances, $\kappa : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_+$. The score for any pair of instances $(i, j)$ is $\kappa(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) \geq 0$, with $\kappa(i, j)$ being large when instances $\boldsymbol{x}^{(i)}$ and $\boldsymbol{x}^{(j)}$ are similar. If the function $\kappa$ obeys a few key properties it is called a **kernel function**.[4]

Given a valid kernel function, we can build a non-linear classifier without explicitly defining a feature vector. For a binary classification problem $y \in \{-1, 1\}$, we have the

[4]The **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 ($\forall \boldsymbol{a}, \boldsymbol{a}^\top \mathbf{K} \boldsymbol{a} \geq 0$). For more on kernel-based classification, see chapter 14 of Murphy (2012).

| | | |
|---|---|---|
| 1. | ***George Bush*** *traveled to* ***France*** | *George Bush* $\underset{\text{NSUBJ}}{\leftarrow}$ *traveled* $\underset{\text{OBL}}{\rightarrow}$ *France* |
| 2. | ***Ahab*** *traveled to* ***Nantucket*** | *Ahab* $\underset{\text{NSUBJ}}{\leftarrow}$ *traveled* $\underset{\text{OBL}}{\rightarrow}$ *Nantucket* |
| 3. | ***George Bush*** *will travel to* ***France*** | *George Bush* $\underset{\text{NSUBJ}}{\leftarrow}$ *travel* $\underset{\text{OBL}}{\rightarrow}$ *France* |
| 4. | ***George Bush*** *wants to travel to* ***France*** | *George Bush* $\underset{\text{NSUBJ}}{\leftarrow}$ *wants* $\underset{\text{XCOMP}}{\rightarrow}$ *travel* $\underset{\text{OBL}}{\rightarrow}$ *France* |
| 5. | ***Ahab*** *traveled to a city in* ***France*** | *Ahab* $\underset{\text{NSUBJ}}{\leftarrow}$ *traveled* $\underset{\text{OBL}}{\rightarrow}$ *city* $\underset{\text{NMOD}}{\rightarrow}$ *France* |
| 6. | *We await* ***Ahab*** *'s visit to France* | *Ahab* $\underset{\text{NMOD:POSS}}{\leftarrow}$ *visit* $\underset{\text{NMOD}}{\rightarrow}$ *France* |

Table 16.2: Candidates instances for the PHYSICAL.LOCATED relation, and their dependency paths

decision function,

$$\hat{y} = \text{Sign}(b + \sum_{i=1}^{N} y^{(i)} \alpha^{(i)} \kappa(\boldsymbol{x}^{(i)}, \boldsymbol{x})) \quad [16.12]$$

where $b$ and $\{\alpha^{(i)}\}_{i=1}^{N}$ are parameters that must be learned from the training set, under the constraint $\forall_i, \alpha^{(i)} \geq 0$. Intuitively, each $\alpha_i$ specifies the importance of the instance $\boldsymbol{x}^{(i)}$ towards the classification rule. Kernel-based classification can be viewed as a weighted form of the **nearest-neighbor** classifier (Hastie et al., 2009), in which test instances are assigned the most common label among their near neighbors in the training set. This results in a non-linear classification boundary. The parameters are typically learned from a margin-based objective (see § 1.3), leading to the **kernel support vector machine**. To generalize to multi-class classification, we can train separate binary classifiers for each label (sometimes called **one-versus-all**), or train binary classifiers for each pair of possible labels (**one-versus-one**).

Dependency kernels are particularly effective for relation extraction, due to their ability to capture syntactic properties of the path between the two candidate arguments. One class of dependency tree kernels is defined recursively, with the score for a pair of trees 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** were an early neural architecture for relation extraction (Zeng et al., 2014; dos Santos et al., 2015). For the sentence $(w_1, w_2, \ldots, w_M)$, obtain a matrix of word embeddings $\mathbf{X}$, where $\boldsymbol{x}_m \in \mathbb{R}^K$ 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), we consider a restricted version of the relation extraction task in which the arguments are single tokens.) We can augment the word embeddings by estimating embeddings for these positional offsets, $\boldsymbol{x}^{(p)}_{m-a_1}$ and $\boldsymbol{x}^{(p)}_{m-a_2}$. The complete base representation of the sentence is,

$$\mathbf{X}(a_1, a_2) = \begin{pmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_M \\ \boldsymbol{x}^{(p)}_{1-a_1} & \boldsymbol{x}^{(p)}_{2-a_1} & \cdots & \boldsymbol{x}^{(p)}_{M-a_1} \\ \boldsymbol{x}^{(p)}_{1-a_2} & \boldsymbol{x}^{(p)}_{2-a_2} & \cdots & \boldsymbol{x}^{(p)}_{M-a_2} \end{pmatrix}, \tag{16.13}$$

where each column is a vertical concatenation of a word embedding, represented by the column vector $\boldsymbol{x}_m$, and two positional embeddings, 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, and max-pooling is applied to obtain a vector. The final scoring function is then,

$$\psi(r, i, j, \mathbf{X}) = \boldsymbol{\theta}_r \cdot \text{MaxPool}(\text{ConvNet}(\mathbf{X}(i, j); \boldsymbol{\phi})), \tag{16.14}$$

where $\boldsymbol{\phi}$ defines the parameters of the convolutional operator, and the $\boldsymbol{\theta}_r$ defines a set of weights for relation $r$. The model can be trained using a margin objective,

$$\hat{r} = \underset{r}{\text{argmax}}\, \psi(r, i, j, \mathbf{X}) \tag{16.15}$$

$$\ell = (1 + \psi(\hat{r}, i, j, \mathbf{X}) - \psi(r, i, j, \mathbf{X}))_+. \tag{16.16}$$

**Recurrent neural networks** have also been applied to relation extraction, using a network such as an 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,

(16.10) *George Bush* $\underset{\text{NSUBJ}}{\leftarrow}$ *wants* $\underset{\text{XCOMP}}{\rightarrow}$ *travel* $\underset{\text{OBL}}{\rightarrow}$ *France*

is segmented into the subpaths,

(16.11) *George Bush* $\underset{\text{NSUBJ}}{\leftarrow}$ *wants*

(16.12) *wants* $\underset{\text{XCOMP}}{\rightarrow}$ *travel* $\underset{\text{OBL}}{\rightarrow}$ *France*.

They then run recurrent networks from the arguments to the root word (in this case, *wants*), obtaining the final representation by max pooling 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. 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 $\boldsymbol{x}_m^{(c)}$ indicate the embedding of word (or relation) $m$ in channel $c$, and let $\overleftarrow{\boldsymbol{h}}_m^{(c)}$ indicate the associated recurrent state in the left subtree. Then the complete model is specified as follows,

$$\boldsymbol{h}_{s(m)}^{(c)} = \text{RNN}(\boldsymbol{x}_{s(m)}^{(c)}, \boldsymbol{h}_m^{(c)}) \quad [16.17]$$

$$\boldsymbol{z}^{(c)} = \text{MaxPool}\left(\overleftarrow{\boldsymbol{h}}_i^{(c)}, \overleftarrow{\boldsymbol{h}}_{s(i)}^{(c)}, \ldots, \overleftarrow{\boldsymbol{h}}_{\text{root}}^{(c)}, \overrightarrow{\boldsymbol{h}}_j^{(c)}, \overrightarrow{\boldsymbol{h}}_{s(j)}^{(c)}, \ldots, \overrightarrow{\boldsymbol{h}}_{\text{root}}^{(c)}\right) \quad [16.18]$$

$$\psi(r, i, j) = \boldsymbol{\theta} \cdot \left[\boldsymbol{z}^{(\text{word})}; \boldsymbol{z}^{(\text{POS})}; \boldsymbol{z}^{(\text{dependency})}; \boldsymbol{z}^{(\text{hypernym})}\right]. \quad [16.19]$$

Note that $\boldsymbol{z}$ is computed by applying max-pooling to the *matrix* of horizontally concatenated vectors $\boldsymbol{h}$, while $\psi$ is computed from the *vector* of vertically concatenated vectors $\boldsymbol{z}$. Xu et al. (2015) pass the score $\psi$ through a **softmax** layer to obtain a probability $\text{p}(r \mid i, j, \boldsymbol{w})$, and train the model by regularized **cross-entropy**. Miwa and Bansal (2016) show that a related model can solve the more challenging "end-to-end" relation extraction task, in which the model must simultaneously detect entities and then extract their relations.

### 16.2.3 Knowledge base population

In many applications, what matters is not what fraction of sentences are analyzed correctly, but how much accurate knowledge can be extracted. **Knowledge base population (KBP)** refers to the task of filling in Wikipedia-style infoboxes, as shown in Figure 16.1a. Knowledge base population can be decomposed into two subtasks: **entity linking** (described in § 16.1), and **slot filling** (Ji and Grishman, 2011). Slot filling has two key differences from the formulation of relation extraction presented above:

- The relations hold between entities, rather than spans of text
- Performance is evaluated at the *type level* (on entity pairs), rather than on the *token level* (on individual sentences).

From a practical standpoint, there are three other important differences between slot filling and per-sentence relation extraction.

- KBP tasks are often formulated from the perspective of identifying attributes of a few "query" entities. As a result, these systems often start with an **information retrieval** phase, in which relevant passages of text are obtained by search.
- For many entity pairs, there will be multiple passages of text that provide evidence. Aggregating this evidence to predict a single relation type (or set of relations) is a challenge for slot filling systems.
- Labeled data is usually available in the form of pairs of related entities, rather than annotated passages of text. Training from such type-level annotations is relatively complex: two entities may be linked by several relations, or they may appear together in a passage of text that nonetheless does not describe their relation to each other.

Information retrieval is beyond the scope of this text (see Manning et al., 2008). The remainder of this section describes approaches to information fusion and learning from type-level annotations.

#### Information fusion

In knowledge base population, there will often be multiple pieces of evidence for (and sometimes against) a single relation. For example, a search for the entity `Maynard Jackson, Jr.` may return several passages that reference the entity `Atlanta`:

(16.13) Elected mayor of **Atlanta** in 1973, **Maynard Jackson** was the first African American to serve as mayor of a major southern city.

(16.14) **Atlanta**'s airport will be renamed to honor **Maynard Jackson**, the city's first Black mayor .

(16.15) Born in Dallas, Texas in 1938, **Maynard Holbrook Jackson, Jr.** moved to **Atlanta** when he was 8.

(16.16) **Maynard Jackson** has gone from one of the worst high schools in **Atlanta** to one of the best.

The first and second examples provide evidence for the relation MAYOR holding between the entities `Atlanta` and `Maynard Jackson, Jr.`. The third example provides evidence for a different relation between these same entities, LIVED-IN. The fourth example poses an entity linking problem, referring to `Maynard Jackson high school`. Knowledge base population requires aggregating this sort of textual evidence, and predicting the relations that are most likely to hold.

One approach is to run a single-document relation extraction system (using the techniques described in § 16.2.2), and then aggregate the results (Li et al., 2011). Relations

that are detected with high confidence in multiple documents are more likely to be valid, motivating the heuristic,

$$\psi(r, e_1, e_2) = \sum_{i=1}^{N} (\text{p}(r(e_1, e_2) \mid \boldsymbol{w}^{(i)}))^{\alpha}, \quad [16.20]$$

where $\text{p}(r(e_1, e_2) \mid \boldsymbol{w}^{(i)})$ is the probability of relation $r$ between entities $e_1$ and $e_2$ conditioned on the text $\boldsymbol{w}^{(i)}$, and $\alpha \gg 1$ is a tunable hyperparameter. Using this heuristic, it is possible to rank all candidate relations, and trace out a **precision-recall curve** as more relations are extracted.[5] Alternatively, features can be aggregated across multiple passages of text, feeding a single type-level relation extraction system (Wolfe et al., 2017).

Precision can be improved by introducing constraints across multiple relations. For example, if we are certain of the relation $\text{PARENT}(e_1, e_2)$, then it cannot also be the case that $\text{PARENT}(e_2, e_1)$. Integer linear programming makes it possible to incorporate such constraints into a global optimization (Li et al., 2011). Other pairs of relations have positive correlations, such $\text{MAYOR}(e_1, e_2)$ and $\text{LIVED-IN}(e_1, e_2)$. Compatibility across relation types can be incorporated into probabilistic graphical models (e.g., Riedel et al., 2010).

#### Distant supervision

Relation extraction is "annotation hungry," because each relation requires its own labeled data. Rather than relying on annotations of individual documents, it would be preferable to use existing knowledge resources — such as the many facts that are already captured in knowledge bases like DBPedia. However such annotations raise the inverse of the information fusion problem considered above: the existence of the relation $\text{MAYOR}(\texttt{Maynard Jackson Jr.}, \texttt{Atlanta})$ provides only **distant supervision** for the example texts in which this entity pair is mentioned.

One approach is to treat the entity pair as the instance, rather than the text itself (Mintz et al., 2009). Features are then aggregated across all sentences in which both entities are mentioned, and labels correspond to the relation (if any) between the entities in a knowledge base, such as FreeBase. Negative instances are constructed from entity pairs that are not related in the knowledge base. In some cases, two entities are related, but the knowledge base is missing the relation; however, because the number of possible entity pairs is huge, these missing relations are presumed to be relatively rare. This approach is shown in Figure 16.2.

In **multiple instance learning**, sets of instances receive a single label, which applies only to an unknown subset (Dietterich et al., 1997; Maron and Lozano-Pérez, 1998). This formalizes the framework of distant supervision: the relation $\text{REL}(\texttt{a}, \texttt{b})$ can act as a label

[5] The precision-recall curve is similar to the ROC curve shown in Figure 3.4, but it includes the precision $\frac{\text{TP}}{\text{TP+FP}}$ rather than the false positive rate $\frac{\text{FP}}{\text{FP+TN}}$.

- **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 16.2: 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.

for the entire set of sentences mentioning entities a and b, even when only a subset of these sentences actually describes the relation. One approach to multi-instance learning is to introduce a binary **latent variable** for each sentence, indicating whether the sentence expresses the labeled relation (Riedel et al., 2010). A variety of inference techniques have been employed for this probabilistic model of relation extraction: Surdeanu et al. (2012) use expectation maximization, Riedel et al. (2010) use sampling, and Hoffmann et al. (2011) use a custom graph-based algorithm. Expectation maximization and sampling are surveyed in chapter 4, and are covered in more detail by Murphy (2012); graph-based methods are surveyed by Mihalcea and Radev (2011).

### 16.2.4 Open information extraction

In classical relation extraction, the set of relations is defined in advance, using a **schema**. The relation for any pair of entities can then be predicted using multi-class classification. In **open information extraction**, a relation can be any triple of text. The example sentence

| 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 16.3: Various relation extraction tasks and their properties. VerbNet and FrameNet are described in chapter 12. 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.

(16.13) instantiates several "relations" of this sort:

- $\langle$*mayor of*, *Maynard Jackson*, *Atlanta*$\rangle$,
- $\langle$*elected*, *Maynard Jackson*, *mayor of Atlanta*$\rangle$,
- $\langle$*elected in*, *Maynard Jackson*, *1973*$\rangle$,

and so on. Extracting such tuples can be viewed as a lightweight version of **semantic role labeling** (chapter 12), with only two argument types: first slot and second slot. The task is generally evaluated on the relation level, rather than on the level of sentences: precision is measured by the number of extracted relations that are accurate, and recall is measured by the number of true relations that were successfully extracted. OpenIE systems are trained from distant supervision or bootstrapping, rather than from labeled sentences.

An early example is the `TextRunner` system (Banko et al., 2007), which identifies relations with a set of handcrafted syntactic rules. The examples that are acquired from the handcrafted rules are then used to train a speedier classification model that uses part-of-speech patterns as features. Finally, the relations that are extracted by the classifier are aggregated, removing redundant relations and computing the number of times that each relation is mentioned in the corpus. `TextRunner` was the first in a series of systems that performed increasingly accurate open relation extraction by incorporating more precise linguistic features (Etzioni et al., 2011), distant supervision from Wikipedia infoboxes (Wu and Weld, 2010), and better learning algorithms (Zhu et al., 2009).

## 16.3 Events

Relations link pairs of entities, but many real-world situations involve more than two entities. Consider again the example sentence (16.13), which describes the **event** of an election, with four properties: the office (`mayor`), the district (`Atlanta`), the date (`1973`), and

the person elected (`Maynard Jackson, Jr.`). In **event detection**, a schema is provided for each event type (e.g., an election, a terrorist attack, or a chemical reaction), indicating all the possible properties of the event. The system is then required to fill in as many of these properties as possible (Doddington et al., 2004).

Event detection systems generally involve a retrieval component (finding relevant documents and passages of text) and an extraction component (determining the properties of the event based on the retrieved texts). Early approaches focused on finite-state patterns for identify event properties (Hobbs et al., 1997); such patterns can be automatically induced by searching for patterns that are especially likely to appear in documents that match the event query (Riloff, 1996). Contemporary approaches employ techniques that are similar to FrameNet semantic role labeling (§ 12.2), such as structured prediction over local and global features (Li et al., 2013) and bidirectional recurrent neural networks (Feng et al., 2016). These methods detect whether an event is described in a sentence, and if so, what are its properties.

**Event coreference** Because multiple sentences may describe unique properties of a single event, **event coreference** is required to link event mentions across a single passage of text, or between passages (Humphreys et al., 1997). Bejan and Harabagiu (2014) define event coreference as the task of identifying event mentions that share the same event participants (i.e., the slot-filling entities) and the same event properties (e.g., the time and location), within or across documents. Event coreference resolution can be performed using supervised learning techniques in a similar way to entity coreference, as described in chapter 14: move left-to-right through the document, and use a classifier to decide whether to link each event reference to an existing cluster of coreferent events, or to create a new cluster (Ahn, 2006). Each clustering decision is based on the compatibility of features describing the participants and properties of the event. Due to the difficulty of annotating large amounts of data for entity coreference, unsupervised approaches are especially desirable (Chen and Ji, 2009; Bejan and Harabagiu, 2014). [todo: figure with example]

**Relations between events** Just as entities are related to other entities, events may be related to other events: for example, the event of winning an election both *precedes* and *causes* the event of serving as mayor; moving to Atlanta *precedes* and *enables* the event of becoming mayor of Atlanta; moving from Dallas to Atlanta *prevents* the event of later becoming mayor of Dallas. As these examples show, events may be related both temporally and causally. The **TimeML** annotation scheme specifies a set of six temporal relations between events (Pustejovsky et al., 2005), derived in part from **interval algebra** (Allen, 1984). The `TimeBank` corpus provides TimeML annotations for 186 documents (Pustejovsky et al., 2003). Methods for detecting these temporal relations have combined supervised machine learning with temporal constraints, such as transitivity (Mani et al., 2006;

| | 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 16.4: Table of factuality values from the FACTBANK corpus (Saurí and Pustejovsky, 2009). The entry (NA) indicates that this combination is not annotated.

Chambers and Jurafsky, 2008).

More recent annotation schemes and datasets have attempted to combine temporal and causal relations (Mirza et al., 2014; Dunietz et al., 2017): for example, the CaTeRS dataset includes annotations of 320 five-sentence short stories (Mostafazadeh et al., 2016). Abstracting still further, **processes** are networks of causal relations between multiple events. A small dataset of biological processes is annotated in the `ProcessBank` dataset (Berant et al., 2014), with the goal of supporting automatic question answering on scientific textbooks.

## 16.4 Hedges, denials, and hypotheticals

The methods described thus far apply to **propositions** about the way things are in the real world. But natural language can also describe events and relations that are likely or unlikely, possible or impossible, desired or feared. The following examples hint at the scope of the problem (Prabhakaran et al., 2010):

(16.17) GM will lay off workers.

(16.18) A spokesman for GM said GM will lay off workers.

(16.19) GM may lay off workers.

(16.20) The politician claimed that GM will lay off workers.

(16.21) Some wish GM would lay off workers.

(16.22) Will GM lay off workers?

(16.23) Many wonder whether GM will lay off workers.

Accurate information extraction requires handling these **extra-propositional** aspects of meaning, which are sometimes summarized under the terms **modality** and **negation**.[6]

[6]The 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

Modality refers to expressions of the speaker's attitude towards her own statements, including "degree of certainty, reliability, subjectivity, sources of information, and perspective" (Morante and Sporleder, 2012). Various systematizations of modality have been proposed (e.g., Palmer, 2001), including categories such as future, interrogative, imperative, conditional, and subjective. Information extraction is particularly concerned with negation and certainty. For example, Saurí and Pustejovsky (2009) link negation with a modal calculus of certainty, likelihood, and possibility, creating the two-dimensional analysis shown in Table 16.4. This schema is the basis for the FACTBANK corpus, with annotations of the **factuality** of all sentences in 208 documents of news text.

A related concept is **hedging**, in which speakers limit their commitment to a proposition (Lakoff, 1973):

(16.24) These results **suggest** that expression of c-jun, jun B and jun D genes **might** be involved in terminal granulocyte differentiation... (Morante and Daelemans, 2009)

(16.25) A whale is **technically** a mammal (Lakoff, 1973)

In the first example, the hedges *suggest* and *might* communicate uncertainty; in the second example, there is no uncertainty, but the hedge *technically* indicates that the evidence for the proposition will not fully meet the reader's expectations. Hedging has been studied extensively in scientific texts (Medlock and Briscoe, 2007; Morante and Daelemans, 2009), where the goal of large-scale extraction of scientific facts is obstructed by hedges and speculation. Still another related aspect of modality is **evidentiality**, in which speakers mark the source of their information. In many languages, it is obligatory to mark evidentiality through affixes or particles (Aikhenvald, 2004); while evidentiality is not grammaticalized in English, authors are expected to express this information in contexts such as journalism (Kovach and Rosenstiel, 2014) and Wikipedia.[7]

Methods for handling negation and modality generally include two phases:

1. detecting negated or uncertain events;

2. identifying the scope and focus of the negation or modal operator.

A considerable body of work on negation has employed rule-based techniques such as regular expressions (Chapman et al., 2001) to detect negated events. Such techniques

of the semantic parsing tasks discussed in chapter 11 and chapter 12 — for example, negation markers are 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.

[7] 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*). More recent approaches employ classifiers over lexical and syntactic features (Uzuner et al., 2009) and sequence labeling (Prabhakaran et al., 2010).

The tasks of scope and focus resolution is more fine grained, as shown in the following example from Morante and Sporleder (2012):

(16.26) [ 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* (this is the focus of negation), and each occurs within a scope: *after his habit he said* and *after mine I asked _____ questions*. These tasks are typically formalized as sequence labeling problems, with each word token labeled as beginning, inside, or outside of a cue, focus, or scope span (see § 7.3). Conventional sequence labeling approaches can then be applied, using surface features as well as syntax (Velldal 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).

## 16.5 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 in the history of natural language processing (Ferrucci et al., 2010). Game show questions are usually answered by **factoids**: entity names and short phrases.[8] The task of factoid question answering is therefore closely related to information extraction, with the additional problem of accurately parsing the question.

### 16.5.1 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?*

[8] The 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).

would be converted to,

$$\lambda x.\exists y \text{ CAPITAL}(\texttt{Georgia}, y) \wedge \text{MAYOR}(y, x). \quad [16.21]$$

This lambda expression can then be used to query an existing knowledge base, returning all entities that satisfy it. The knowledge base itself could be constructed using techniques described in § 16.2.3.

### 16.5.2 Machine reading

Recent work has focused on answering questions about specific textual passages, similar to the reading comprehension examinations for young students (Hirschman et al., 1999). This task has come to be known as **machine reading**.

#### Datasets

The machine reading problem can be formulated in a number of different ways. The most important distinction is what form the answer should take.

- **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, the answer is deducible from the text alone, while in the science exams, the system must make inferences using an existing model of the underlying scientific phenomena. Here is an example from MCTEST:

  (16.27) James the turtle was always getting into trouble. Sometimes he'd reach into the freezer and empty out all the food ...
  Q: What is the name of the trouble making turtle?
  (a) Fries
  (b) Pudding
  (c) James
  (d) Jane

- **Cloze**-style "fill in the blank" questions, as in the CNN/DAILY MAIL comprehension task (Hermann et al., 2015), the CHILDREN'S BOOK TEST (Hill et al., 2016), and the WHO-DID-WHAT dataset (Onishi et al., 2016). In these tasks, the system must guess which word or entity completes a sentence, based on reading a passage of text. Here is an example from WHO-DID-WHAT:

  (16.28) Q: Tottenham manager Juande Ramos has hinted he will allow _____ to leave if the Bulgaria striker makes it clear he is unhappy. (Onishi et al., 2016)

The query sentence may be selected either from the story itself, or from an external summary. In either case, datasets can be created automatically by processing large quantities existing documents. An additional constraint is that that missing element from the cloze must appear in the main passage of text: for example, in WHO-DID-WHAT, the candidates include all entities mentioned in the main passage. In the CNN/DAILY MAIL dataset, each entity name is replaced by a unique identifier, e.g., `entity37`. This ensures that correct answers can only be obtained by accurately reading the text, and not from external knowledge about the entities.

- **Extractive** question answering, in which the answer is drawn from the original text. In WIKIQA, answers are sentences (Yang et al., 2015); in the Stanford Question Answering Dataset (SQUAD), answers are words or short phrases (Rajpurkar et al., 2016):

  (16.29) In metereology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity.
  Q: What causes precipitation to fall? A: gravity

  In both WIKIQA and SQUAD, the original texts are Wikipedia articles, and the questions are generated by crowdworkers.

### Methods

A baseline method is to search the text for sentences or short passages that overlap with both the query and the candidate answer (Richardson et al., 2013). In example (16.27), this baseline would select the correct answer (c), since *James* appears in a sentence that includes the query terms *trouble* and *turtle*.

This baseline can be implemented as a neural architecture, using an **attention mechanism** that 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 $\boldsymbol{w}^{(p)}$ and the query $\boldsymbol{w}^{(q)}$, using two bidirectional LSTMs (§ 6.5.4).

$$\boldsymbol{h}^{(q)} = \text{BiLSTM}(\boldsymbol{w}^{(q)}; \Theta^{(q)}) \quad [16.22]$$

$$\boldsymbol{h}^{(p)} = \text{BiLSTM}(\boldsymbol{w}^{(p)}; \Theta^{(p)}). \quad [16.23]$$

The query is represented by vertically concatenating the final states of the left-to-right and right-to-left passes:

$$\boldsymbol{u} = [\overrightarrow{\boldsymbol{h}^{(q)}}_{M_q}; \overleftarrow{\boldsymbol{h}^{(q)}}_{0}]. \quad [16.24]$$

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,

$$\tilde{\alpha}_m = (\boldsymbol{u}^{(q)})^\top \mathbf{W}_a \boldsymbol{h}_m^{(p)} \quad [16.25]$$

$$\boldsymbol{\alpha} = \text{SoftMax}(\tilde{\boldsymbol{\alpha}}) \quad [16.26]$$

$$\boldsymbol{o} = \sum_{m=1}^{M} \alpha_m \boldsymbol{h}_m^{(p)}. \quad [16.27]$$

Each candidate answer $c$ is represented by a vector $\boldsymbol{x}_c$. Assuming the candidate answers are spans from the original text, these vectors can be set equal to the corresponding element in $\boldsymbol{h}^{(p)}$. The score for each candidate answer $a$ is computed by the inner product,

$$\hat{c} = \underset{c}{\text{argmax}}\, \boldsymbol{o} \cdot \boldsymbol{x}_c. \quad [16.28]$$

This architecture can be trained end-to-end from a loss based on the log-likelihood of the correct answer. A number of related architectures have been proposed (e.g., Hermann et al., 2015; Kadlec et al., 2016; Dhingra et al., 2017; Cui et al., 2017), and the relationships between these methods are surveyed by Wang et al. (2017).

## Additional reading

The field of information extraction is surveyed in course notes by Grishman (2012), and more recently in a short survey paper (Grishman, 2015). Shen et al. (2015) survey the task of entity linking, and Ji and Grishman (2011) survey work on knowledge base population. This chapter's discussion of non-propositional meaning was strongly influenced by Morante and Sporleder (2012), who introduced a special issue of the journal *Computational Linguistics* dedicated to recent work on modality and negation.

## Exercises

1. Consider the following heuristic for entity linking:
   - Among all entities that have the same type as the mention (e.g., LOC, PER), choose the one whose name has the lowest edit distance from the mention.
   - If more than one entity has the right type and the lowest edit distance from the mention, choose the most popular one.
   - If no candidate entity has the right type, choose NIL.

   Now suppose you have the following feature function:

   $$\boldsymbol{f}(y, \boldsymbol{x}) = [\text{edit-dist}(\text{name}(y), \boldsymbol{x}), \text{same-type}(y, \boldsymbol{x}), \text{popularity}(y), \delta(y = \text{NIL})]$$

Design a set of ranking weights $\boldsymbol{\theta}$ that match the heuristic. You may assume that edit distance and popularity are always in the range $[0, 100]$, and that the NIL entity has values of zero for all features except $\delta(y = \text{NIL})$.

2. Now consider another heuristic:

   - Among all candidate entities that have edit distance zero from the mention and the right type, choose the most popular one.
   - If no entity has edit distance zero from the mention, choose the one with the right type that is most popular, regardless of edit distance.
   - If no entity has the right type, choose NIL.

   Using the same features and assumptions from the previous problem, prove that there is no set of weights that could implement this heuristic. Then show that the heuristic can be implemented by adding a single feature. Your new feature should consider only the edit distance.

3. * Consider the following formulation for collective entity linking, which rewards sets of entities that are all of the same type, where "types" can be elements of any set:

$$\psi_c(\boldsymbol{y}) = \begin{cases} \alpha & \text{all entities in } \boldsymbol{y} \text{ have the same type} \\ \beta & \text{more than half of the entities in } \boldsymbol{y} \text{ have the same type} \\ 0 & \text{otherwise.} \end{cases} \quad [16.29]$$

   Show how to implement this model of collective entity linking in a **integer linear program**. You may want to review § 12.2.2.

   To get started, here is an integer linear program for entity linking, without including the collective term $\psi_c$:

$$\max_{z_{i,y} \in \{0,1\}} \sum_{i=1}^{N} \sum_{y \in \mathcal{Y}(\boldsymbol{x}^{(i)})} s_{i,y} z_{i,y}$$
$$\text{s.t.} \sum_{y \in \mathcal{Y}(x^{(i)})} z_{i,y} \leq 1 \quad \forall i \in \{1, 2, \ldots N\}$$

   where $z_{i,y} = 1$ if entity $y$ is linked to mention $i$, and $s_{i,y}$ is a parameter that scores the quality of this individual ranking decision, e.g., $s_{i,y} = \boldsymbol{\theta} \cdot \boldsymbol{f}(y, \boldsymbol{x}^{(i)}, \boldsymbol{c}^{(i)})$.

   To incorporate the collective linking score, you may assume parameters $r$,

$$r_{y,\tau} = \begin{cases} 1, & \text{entity } y \text{ has type } \tau \\ 0, & \text{otherwise.} \end{cases} \quad [16.30]$$

   **Hint**: You will need to define several auxiliary variables to optimize over.

4. Run `nltk.corpus.download('reuters')` to download the Reuters corpus in NLTK, and run `from nltk.corpus import reuters` to import it. The command `reuters.words()` returns an iterator over the tokens in the corpus.

   a) Apply the pattern _____, *such as* _____ to this corpus, obtaining candidates for the IS-A relation, e.g. IS-A(`Romania`, `Country`). What are three pairs that this method identifies correctly? What are three different pairs that it gets wrong?

   b) Design a pattern for the PRESIDENT relation, e.g. PRESIDENT(`Philippines`, `Corazon Aquino`). In this case, you may want to augment your pattern matcher with the ability to match multiple token wildcards, perhaps using case information to detect proper names. Again, list three correct

   c) Preprocess the Reuters data by running a named entity recognizer, replacing tokens with named entity spans when applicable. Apply your PRESIDENT matcher to this new data. Does the accuracy improve? Compare 20 randomly-selected pairs from this pattern and the one you designed in the previous part.

5. Represent the dependency path $\boldsymbol{x}^{(i)}$ as a sequence of words and dependency arcs of length $M_i$, ignoring the endpoints of the path. In example 1 of Table 16.2, the dependency path is,

$$\boldsymbol{x}^{(1)} = (\underset{\text{NSUBJ}}{\leftarrow}, \textit{traveled}, \underset{\text{OBL}}{\rightarrow}) \quad [16.31]$$

   If $x_m^{(i)}$ is a word, then let $\text{pos}(x_m^{(i)})$ be its part-of-speech, using the tagset defined in chapter 7.

   We can define the following kernel function over pairs of dependency paths (Bunescu and Mooney, 2005):

$$\kappa(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) = \begin{cases} 0, & M_i \neq M_j \\ \prod_{m=1}^{M_i} c(x_m^{(i)}, x_m^{(j)}), & M_i = M_j \end{cases}$$

$$c(x_m^{(i)}, x_m^{(j)}) = \begin{cases} 2, & x_m^{(i)} = x_m^{(j)} \\ 1, & x_m^{(i)} \text{ and } x_m^{(j)} \text{ are words and } \text{pos}(x_m^{(i)}) = \text{pos}(x_m^{(j)}) \\ 0, & \text{otherwise.} \end{cases}$$

   Using this kernel function, compute the kernel similarities of example 1 from Table 16.2 with the other five examples.

6. Continuing from the previous problem, suppose that the instances have the following labels:

$$y_2 = 1, y_3 = -1, y_4 = -1, y_5 = 1, y_6 = 1 \quad [16.32]$$

   Identify the conditions for $\alpha$ and $b$ under which $\hat{y}_1 = 1$. Remember the constraint that $\alpha_i \geq 0$ for all $i$.

# Chapter 17

# Text generation

# Chapter 18

# Machine translation

# Appendix A

# Probability

Probability theory provides a way to reason about random events. The sorts of random events that are typically used to explain probability theory include coin flips, card draws, and the weather. It may seem odd to think about the choice of a word as akin to the flip of a coin, particularly if you are the type of person to choose words carefully. But random or not, language has proven to be extremely difficult to model deterministically. Probability offers a powerful tool for modeling and manipulating linguistic data, which we will use repeatedly throughout this course.

Probability can be thought of in terms of **random outcomes**: for example, a single coin flip has two possible outcomes, heads or tails. The set of possible outcomes is the **sample space**, and a subset of the **sample space** is an **event**. For a sequence of two coin flips, there are four possible outcomes, $\{HH, HT, TH, TT\}$, representing the ordered sequences heads-head, heads-tails, tails-heads, and tails-tails. The event of getting exactly one head includes two outcomes: $\{HT, TH\}$.

Formally, a probability is a function from events to the interval between zero and one: $\Pr : \mathcal{F} \rightarrow [0, 1]$, where $\mathcal{F}$ is the set of possible events. An event that is certain has probability one; an event that is impossible has probability zero. For example, the probability of getting less than three heads on two coin flips is one. Each outcome is also an event (a set with exactly one element), and for two flips of a fair coin, the probability of each outcome is,

$$\Pr(\{HH\}) = \Pr(\{HT\}) = \Pr(\{TH\}) = \Pr(\{TT\}) = \frac{1}{4}. \quad \text{[A.1]}$$

## A.1 Probabilities of event combinations

Because events are sets of outcomes, we can use set-theoretic operations such as complement, intersection, and unions to reason about the probabilities of various event combinations.

For any event $A$, there is a **complement** $\neg A$, such that:

- The union $A \cup \neg A$ covers the entire sample space, and $\Pr(A \cup \neg A) = 1$;
- The intersection $A \cap \neg A = \varnothing$ is the empty set, and $\Pr(A \cap \neg A) = 0$.

In the coin flip example, the event of obtaining a single head on two flips corresponds to the set of outcomes $\{HT, TH\}$; the complement event includes the other two outcomes, $\{TT, HH\}$.

### A.1.1 Probabilities of disjoint events

In general, when two events have an empty intersection, $A \cap B = \varnothing$, they are said to be **disjoint**. The probability of the union of two disjoint events is equal to the sum of their probabilities,

$$A \cap B = \varnothing \quad \Rightarrow \quad \Pr(A \cup B) = \Pr(A) + \Pr(B). \tag{A.2}$$

This is the **third axiom of probability**, and can be generalized to any countable sequence of disjoint events.

In the coin flip example, we can use this axiom to 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,

$$\Pr(\{HT, TH\}) = \Pr(\{HT\} \cup \{TH\}) = \Pr(\{HT\}) + \Pr(\{TH\}) \tag{A.3}$$

$$= \frac{1}{4} + \frac{1}{4} = \frac{1}{2}. \tag{A.4}$$

For events that are not disjoint, it is still possible to compute the probability of their union:

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B). \tag{A.5}$$

This can be derived from the third axiom of probability. Consider an event that includes all outcomes in $B$ that are not in $A$, which we can write as $B - (A \cap B)$. By construction, this event is disjoint from $A$.[1] We can therefore apply the additive rule,

$$\Pr(A \cup B) = \Pr(A) + \Pr(B - (A \cap B)) \tag{A.6}$$

$$\Pr(B) = \Pr(B - (A \cap B)) + \Pr(A \cap B) \tag{A.7}$$

$$\Pr(B - (A \cap B)) = \Pr(B) - \Pr(A \cap B) \tag{A.8}$$

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B). \tag{A.9}$$

[todo: simplify]

[1] [todo: add figure]

### A.1.2 Law of total probability

A set of events $\mathcal{B} = \{B_1, B_2, \ldots, B_N\}$ is a **partition** of the sample space iff each pair of events is disjoint ($B_i \cap B_j = \varnothing$), and the union of the events is the entire sample space. The law of total probability states that we can **marginalize** over these events as follows,

$$\Pr(A) = \sum_{B_n \in \mathcal{B}} \Pr(A \cap B_n). \quad [A.10]$$

Note for any event $B$, the union $B \cup \neg B$ forms a partition of the sample space. Therefore, an important special case of the law of total probability is,

$$\Pr(A) = \Pr(A \cap B) + \Pr(A \cap \neg B). \quad [A.11]$$

## 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. We define conditional probability as the ratio,

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)} \quad [A.12]$$

The **chain rule** states that $\Pr(A \cap B) = \Pr(A \mid B) \times \Pr(B)$, which is just a rearrangement of terms from Equation A.12. We can apply the chain rule repeatedly:

$$\begin{aligned}\Pr(A \cap B \cap C) &= \Pr(A \mid B \cap C) \times \Pr(B \cap C) \\ &= \Pr(A \mid B \cap C) \times \Pr(B \mid C) \times \Pr(C)\end{aligned}$$

**Bayes' rule** (sometimes called Bayes' law or Bayes' theorem) gives us a way to convert between $\Pr(A \mid B)$ and $\Pr(B \mid A)$. It follows from the chain rule:

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{\Pr(B \mid A) \times \Pr(A)}{\Pr(B)} \quad [A.13]$$

The terms in Bayes rule have specialized names, which we will occasionally use:

- $\Pr(A)$ is the **prior**, since it is the probability of event $A$ without knowledge about whether $B$ happens or not.

- $\Pr(B \mid A)$ is the **likelihood**, the probability of event $B$ given that event $A$ has occurred.
- $\Pr(A \mid B)$ is the **posterior**, since it is the probability of event $A$ with knowledge that $B$ has occurred.

**Example** Manning and Schütze (1999) have a nice example of Bayes' rule (sometimes called Bayes Law) in a linguistic setting. (This same example is usually framed in terms of tests for rare diseases.) Suppose one is interested in a rare syntactic construction, such as **parasitic gaps**, which occur on average once in 100,000 sentences. Here is an example:

(A.1) *Which class did you attend __ without registering for __?*

Lana Linguist has developed a complicated pattern matcher that attempts to identify sentences with parasitic gaps. It's pretty good, but it's not perfect:

- If a sentence has a parasitic gap, the pattern matcher will find it with probability $0.95$. (This is the **recall**, which is one minus the **false positive rate**.)
- If the sentence doesn't have a parasitic gap, the pattern matcher will wrongly say it does with probability 0.005. (This is the **false positive rate**, which is one minus the **precision**.)

Suppose that Lana's pattern matcher says that a sentence contains a parasitic gap. What 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 $\Pr(G \mid T)$, and we can compute it from Bayes' rule:

$$\Pr(G \mid T) = \frac{\Pr(T \mid G) \times \Pr(G)}{\Pr(T)}. \quad [A.14]$$

We already know both terms in the numerator: $\Pr(T \mid G)$ is the recall, which is $0.95$; $\Pr(G)$ is the prior, which is $10^{-5}$.

We are not given the denominator, but we can compute it by using some of the tools that we have developed in this section. We first apply the law of total probability, using the partition $\{G, \neg G\}$:

$$\Pr(T) = \Pr(T \cap G) + \Pr(T \cap \neg G). \quad [A.15]$$

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** ($T \cap \neg G$). Next, we can

compute the probability of each of these events using the chain rule:

$$\Pr(T \cap G) = \Pr(T \mid G) \times \Pr(G) = 0.95 \times 10^{-5} \quad [A.16]$$
$$\Pr(T \cap \neg G) = \Pr(T \mid \neg G) \times \Pr(\neg G) = 0.005 \times (1 - 10^{-5}) \approx 0.005 \quad [A.17]$$
$$\Pr(T) = \Pr(T \cap G) + \Pr(T \cap \neg G) \quad [A.18]$$
$$= 0.95 \times 10^{-5} + 0.005 \approx 0.005. \quad [A.19]$$

We now return to Bayes' rule to compute the desired posterior probability,

$$\Pr(G \mid T) = \frac{\Pr(T \mid G) \Pr(G)}{\Pr(T)} \quad [A.20]$$
$$= \frac{0.95 \times 10^{-5}}{0.95 \times 10^{-5} + 0.005 \times (1 - 10^{-5})} \quad [A.21]$$
$$\approx 0.002. \quad [A.22]$$

Lana's pattern matcher seems accurate, with false positive and false negative rates below 5%. Yet the extreme rarity of this phenomenon means that a positive result from the detector is most likely to be wrong.

## 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 key role in probabilistic models such as Naïve Bayes chapter 1.

## A.4 Random variables

Random variables are functions of events. Formally, we will treat random variables as functions from events to the space $\mathbb{R}^n$, where $\mathbb{R}$ is the set of real numbers. This general notion subsumes a number of different types of random variables:

- **Indicator random variables** are functions from events to the set $\{0, 1\}$. In the coin flip example, we can define $Y$ as an indicator random variable, for whether the coin has come up heads on at least one flip. This would include the outcomes $\{HH, HT, TH\}$. The event 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 countable subset of $\mathbb{R}$. Consider the coin flip example: the number of heads, $X$, 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 $\Pr(X = 1) = \frac{1}{2}$.

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 $\mathrm{p}_X$, which maps from the possible values of $X$ to the non-negative reals. For a specific value $x$, we write $\mathrm{p}_X(x)$, which is equal to the event probability $\Pr(X = x)$.[2] The function $\mathrm{p}_X$ is called a probability **mass** function (pmf) if $X$ is discrete; it is called a probability **density** function (pdf) if $X$ is continuous. In either case, we have $\int_x \mathrm{p}_X(x)dx = 1$ and $\forall x, \mathrm{p}_X(x) \geq 0$.

Probabilities over multiple Random variables can written as **joint probabilities**, e.g., $\mathrm{p}_{A,B}(a, b) = \Pr(A = a \cap B = b)$. Several ideas from event probabilities carry over to probability distributions over random variables:

- We can write a **marginal probability distribution** $\mathrm{p}_A(a) = \sum_b \mathrm{p}_{A,B}(a, b)$.
- We can write a **conditional probability distribution** $\mathrm{p}_{A|B}(a \mid b) = \frac{\mathrm{p}_{A,B}(a,b)}{\mathrm{p}_B(b)}$.
- Random variables $A$ and $B$ are independent iff $\mathrm{p}_{A,B}(a, b) = \mathrm{p}_A(a) \times \mathrm{p}_B(b)$.

[2] In 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 $\mathrm{p}(x)$.

## A.5 Expectations

Sometimes we want the **expectation** of a function, such as $E[g(x)] = \sum_{x \in \mathcal{X}} g(x)\text{p}(x)$. Expectations are easiest to think about in terms of probability distributions over discrete events:

- If it is sunny, Marvin will eat three ice creams.
- If it is rainy, he 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 sum becomes an integral:

$$E[g(x)] = \int_{\mathcal{X}} g(x)\text{p}(x)dx \quad \text{[A.28]}$$

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 they use a thermometer with infinite precision, the expected price would be an integral over all possible temperatures,

$$E[\text{price}(x)] = \int_{\mathcal{X}} \min(1, 1 + x) \times \text{original-price} \times \text{p}(x)dx. \quad \text{[A.29]}$$

Economically-minded readers will note that the restaurant will apparently pay you for taking poutine, if the temperature falls below $-100$ degrees celsius.

## A.6 Modeling and estimation

**Probabilistic models** give us a principled way to reason about random events and random variables, and to make predictions about the future. Let's consider the coin toss example. We can model each toss as a random event, with probability $\theta$ of the event $H$, and probability $1 - \theta$ of the complementary event $T$. If we write a random variable $X$ as the total number of heads on three coin flips, then the distribution of $X$ depends on $\theta$. In this case, $X$ is distributed as a **binomial random variable**, meaning that it is drawn from a binomial distribution, with **parameters** $(\theta, N = 3)$. We write:

$$X \sim \text{Binomial}(\theta, N = 3). \quad \text{[A.30]}$$

This is a probabilistic model of $X$. The properties of the binomial distribution enable us to make statements about the $X$, such as its expected value and the likelihood that its value will fall within some interval.

Now suppose that $\theta$ is unknown, but we have run an experiment, in which we executed $N$ trials, and obtained $x$ heads. We can **estimate** $\theta$ by the principle of **maximum likelihood**:

$$\hat{\theta} = \underset{\theta}{\text{argmax}}\, \text{p}_X(x; \theta, N). \quad [A.31]$$

This says that our estimate $\hat{\theta}$ should be the value that maximizes the likelihood of the data we have observed. The semicolon indicates that $\theta$ and $N$ are parameters of the probability function. The likelihood $\text{p}_X(x; \theta, N)$ can be computed from the binomial distribution,

$$\text{p}_X(x; \theta, N) = \frac{N!}{x!(N-x)!}\theta^x(1-\theta)^{N-x}. \quad [A.32]$$

This likelihood is proportional to the product of the probability of individual outcomes: for example, the sequence $T, H, H, T, H$ would have probability $\theta^2(1-\theta)^3$. The term $\frac{N!}{x!(N-x)!}$ arises from the many possible orderings by which we could obtain $x$ heads on $N$ trials. This term is constant in $\theta$, so it can be ignored.

We can maximize likelihood by taking the derivative and setting it equal to zero. In practice, we usually maximize log-likelihood, which is a monotonic function of the likelihood, and is easier to manipulate mathematically.

$$\ell(\theta) = x \log \theta + (N-x)\log(1-\theta) \quad [A.33]$$

$$\frac{\partial \ell(\theta)}{\partial \theta} = \frac{x}{\theta} - \frac{N-x}{1-\theta} \quad [A.34]$$

$$\frac{N-x}{1-\theta} = \frac{x}{\theta} \quad [A.35]$$

$$\frac{N-x}{x} = \frac{1-\theta}{\theta} \quad [A.36]$$

$$\frac{N}{x} - 1 = \frac{1}{\theta} - 1 \quad [A.37]$$

$$\hat{\theta} = \frac{x}{N}. \quad [A.38]$$

In this case, the maximum likelihood estimate is equal to $\frac{x}{N}$, the fraction of trials that came up heads. This intuitive solution is also known as the **relative frequency estimate**, 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 $\theta = 1$, so this coin is guaranteed to give heads? If not, then you must have some **prior expectation** about $\theta$. To incorporate this

prior information, we can treat $\theta$ as a random variable, and use Bayes' rule:

$$\begin{aligned} \mathrm{p}(\theta \mid x; N) =& \frac{\mathrm{p}(x \mid \theta) \times \mathrm{p}(\theta)}{\mathrm{p}(x)} && \text{[A.39]} \\ \propto& \mathrm{p}(x \mid \theta) \times \mathrm{p}(\theta) && \text{[A.40]} \\ \hat{\theta} =& \operatorname*{argmax}_{\theta} \mathrm{p}(x \mid \theta) \times \mathrm{p}(\theta). && \text{[A.41]} \end{aligned}$$

This it the **maximum a posteriori** (MAP) estimate. Given a form for $\mathrm{p}(\theta)$, you can derive the MAP estimate using the same approach that was used to derive the maximum likelihood estimate.

## A.7 Further reading

A good introduction to probability theory is offered by Manning and Schütze (1999), which helped to motivate this section. For more detail, Sharon Goldwater provides another useful reference, `http://homepages.inf.ed.ac.uk/sgwater/teaching/general/probability.pdf`. A historical and philosophical perspective on probability is offered by Diaconis and Skyrms (2017).

# Appendix B

# Computational complexity

[todo: 2-3 paragraphs about computational complexity: big-O notation, P vs NP.] The reader is encouraged to consult Arora and Barak (2009) and Sipser (2012) for more on this topic.

# Bibliography


Abadi, M., A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. J. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Józefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. G. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. A. Tucker, V. Vanhoucke, V. Vasudevan, F. B. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *CoRR abs/1603.04467*.

Abend, O. and A. Rappoport (2017). The state of the art in semantic representation. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Abney, S., R. E. Schapire, and Y. Singer (1999). Boosting applied to tagging and PP attachment. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 132–134.

Abney, S. P. (1987). *The English noun phrase in its sentential aspect*. Ph. D. thesis, Massachusetts Institute of Technology.

Abney, S. P. and M. Johnson (1991). Memory requirements and local ambiguities of parsing strategies. *Journal of Psycholinguistic Research 20*(3), 233–250.

Ahn, D. (2006). The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pp. 1–8. Association for Computational Linguistics.

Aikhenvald, A. Y. (2004). *Evidentiality*. Oxford University Press.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19*(6), 716–723.

Akmajian, A., R. A. Demers, A. K. Farmer, and R. M. Harnish (2010). *Linguistics: An introduction to language and communication* (Sixth ed.). Cambridge, MA: MIT press.

Allauzen, C., M. Riley, J. Schalkwyk, W. Skut, and M. Mohri (2007). OpenFst: A general and efficient weighted finite-state transducer library. In *International Conference on Implementation and Application of Automata*, pp. 11–23. Springer.

Allen, J. F. (1984). Towards a general theory of action and time. *Artificial intelligence 23*(2), 123–154.

Alm, C. O., D. Roth, and R. Sproat (2005). Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 579–586.

Aluísio, S., J. Pelizzoni, A. Marchi, L. de Oliveira, R. Manenti, and V. Marquiafável (2003). An account of the challenge of tagging a reference corpus for Brazilian Portuguese. *Computational Processing of the Portuguese Language*, 194–194.

Anand, P., M. Walker, R. Abbott, J. E. Fox Tree, R. Bowmani, and M. Minor (2011). Cats rule and dogs drool!: Classifying stance in online debate. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, Portland, Oregon, pp. 1–9. Association for Computational Linguistics.

Anandkumar, A. and R. Ge (2016). Efficient approaches for escaping higher order saddle points in non-convex optimization. In *Proceedings of the Conference On Learning Theory (COLT)*, pp. 81–102.

Anandkumar, A., R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky (2014). Tensor decompositions for learning latent variable models. *The Journal of Machine Learning Research 15*(1), 2773–2832.

Ando, R. K. and T. Zhang (2005). A framework for learning predictive structures from multiple tasks and unlabeled data. *The Journal of Machine Learning Research 6*, 1817–1853.

Andor, D., C. Alberti, D. Weiss, A. Severyn, A. Presta, K. Ganchev, S. Petrov, and M. Collins (2016). Globally normalized transition-based neural networks. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 2442–2452.

Aronoff, M. (1976). *Word formation in generative grammar*. MIT Press.

Arora, S. and B. Barak (2009). *Computational complexity: a modern approach*. Cambridge University Press.

Arora, S., R. Ge, Y. Halpern, D. Mimno, A. Moitra, D. Sontag, Y. Wu, and M. Zhu (2013). A practical algorithm for topic modeling with provable guarantees. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 280–288.

Arora, S., Y. Li, Y. Liang, T. Ma, and A. Risteski (2016). Linear algebraic structure of word senses, with applications to polysemy. *arXiv preprint arXiv:1601.03764*.

Artstein, R. and M. Poesio (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics 34*(4), 555–596.

Artzi, Y. and L. Zettlemoyer (2013). Weakly supervised learning of semantic parsers for mapping instructions to actions. *Transactions of the Association for Computational Linguistics 1*, 49–62.

Attardi, G. (2006). Experiments with a multilanguage non-projective dependency parser. In *Proceedings of the Conference on Natural Language Learning (CoNLL)*, pp. 166–170.

Auer, P. (2013). *Code-switching in conversation: Language, interaction and identity*. Routledge.

Auer, S., C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives (2007). Dbpedia: A nucleus for a web of open data. *The semantic web*, 722–735.

Austin, J. L. (1962). *How to do things with words*. Oxford University Press.

Aw, A., M. Zhang, J. Xiao, and J. Su (2006). A phrase-based statistical model for SMS text normalization. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 33–40.

Ba, J. L., J. R. Kiros, and G. E. Hinton (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.

Bagga, A. and B. Baldwin (1998a). Algorithms for scoring coreference chains. In *Proceedings of the Language Resources and Evaluation Conference*, pp. 563–566.

Bagga, A. and B. Baldwin (1998b). Entity-based cross-document coreferencing using the vector space model. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pp. 79–85.

Bahdanau, D., K. Cho, and Y. Bengio (2014). Neural machine translation by jointly learning to align and translate. In *Neural Information Processing Systems (NIPS)*.

Baldwin, T. and S. N. Kim (2010). Multiword expressions. In *Handbook of natural language processing*, Volume 2, pp. 267–292. Boca Raton, USA: CRC Press.

Balle, B., A. Quattoni, and X. Carreras (2011). A spectral learning algorithm for finite state transducers. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML)*, pp. 156–171.

Banarescu, L., C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, and N. Schneider (2013, August). Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, Sofia, Bulgaria, pp. 178–186. Association for Computational Linguistics.

Banko, M., M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni (2007). Open information extraction from the web. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 2670–2676.

Bansal, N., A. Blum, and S. Chawla (2004). Correlation clustering. *Machine Learning 56*(1-3), 89–113.

Barber, D. (2012). *Bayesian reasoning and machine learning*. Cambridge University Press.

Barman, U., A. Das, J. Wagner, and J. Foster (2014, October). Code mixing: A challenge for language identification in the language of social media. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, Doha, Qatar, pp. 13–23. Association for Computational Linguistics.

Barnickel, T., J. Weston, R. Collobert, H.-W. Mewes, and V. Stümpflen (2009). Large scale application of neural network based semantic role labeling for automated relation extraction from biomedical texts. *PLoS One 4*(7), e6393.

Baron, A. and P. Rayson (2008). Vard2: A tool for dealing with spelling variation in historical corpora. In *Postgraduate conference in corpus linguistics*.

Baroni, M., R. Bernardi, and R. Zamparelli (2014). Frege in space: A program for compositional distributional semantics. *Linguistic Issues in Language Technologies*.

Baroni, M. and R. Zamparelli (2010). Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 1183–1193.

Beesley, K. R. and L. Karttunen (2003). *Finite-state morphology*. Stanford, CA: Center for the Study of Language and Information.

Bejan, C. A. and S. Harabagiu (2014). Unsupervised event coreference resolution. *Computational Linguistics 40*(2), 311–347.

Bell, E. T. (1934). Exponential numbers. *The American Mathematical Monthly 41*(7), 411–419.

Bender, E. M. (2013, jun). *Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax*, Volume 6 of *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool Publishers.

Bengio, Y., R. Ducharme, P. Vincent, and C. Janvin (2003). A neural probabilistic language model. *The Journal of Machine Learning Research 3*, 1137–1155.

Bengio, Y., P. Simard, and P. Frasconi (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks 5*(2), 157–166.

Bengtson, E. and D. Roth (2008). Understanding the value of features for coreference resolution. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 294–303.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289–300.

Berant, J., A. Chou, R. Frostig, and P. Liang (2013). Semantic parsing on freebase from question-answer pairs. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 1533–1544.

Berant, J., V. Srikumar, P.-C. Chen, A. Vander Linden, B. Harding, B. Huang, P. Clark, and C. D. Manning (2014). Modeling biological processes for reading comprehension. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.

Berg-Kirkpatrick, T., A. Bouchard-Côté, J. DeNero, and D. Klein (2010). Painless unsupervised learning with features. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 582–590.

Berg-Kirkpatrick, T., D. Burkett, and D. Klein (2012). An empirical investigation of statistical significance in NLP. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 995–1005.

Berger, A. L., V. J. D. Pietra, and S. A. D. Pietra (1996). A maximum entropy approach to natural language processing. *Computational linguistics 22*(1), 39–71.

Bergsma, S., D. Lin, and R. Goebel (2008). Distributional identification of non-referential pronouns. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 10–18.

Bergstra, J., O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio (2010). Theano: A CPU and GPU math compiler in Python. In *Proceedings of the 9th Python in Science Conference*, pp. 1–7.

Bertsekas, D. P. (2012). Incremental gradient, subgradient, and proximal methods for convex optimization: A survey. See Sra et al. (2012).

Bhatia, P., R. Guthrie, and J. Eisenstein (2016). Morphological priors for probabilistic neural word embeddings. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.

Biber, D. (1991). *Variation across speech and writing*. Cambridge University Press.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Björkelund, A. and P. Nugues (2011). Exploring lexicalized features for coreference resolution. In *Proceedings of the Conference on Natural Language Learning (CoNLL)*, pp. 45–50.

Blackburn, P. and J. Bos (2005). *Representation and inference for natural language: A first course in computational semantics*. CSLI.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM 55*(4), 77–84.

Blei, D. M. (2014). Build, compute, critique, repeat: Data analysis with latent variable models. *Annual Review of Statistics and Its Application 1*, 203–232.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent dirichlet allocation. *the Journal of machine Learning research 3*, 993–1022.

Blitzer, J., M. Dredze, and F. Pereira (2007). Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 440–447.

Blum, A. and T. Mitchell (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of the Conference On Learning Theory (COLT)*, pp. 92–100.

Böhmová, A., J. Hajič, E. Hajičová, and B. Hladká (2003). The Prague dependency treebank. In *Treebanks*, Text, Speech and Language Technology, pp. 103–127. Springer.

Bohnet, B. (2010). Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pp. 89–97.

Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics 5*, 135–146.

Bollacker, K., C. Evans, P. Paritosh, T. Sturge, and J. Taylor (2008). Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the ACM International Conference on Management of Data (SIGMOD)*, pp. 1247–1250. AcM.

Bordes, A., N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko (2013). Translating embeddings for modeling multi-relational data. In *Neural Information Processing Systems (NIPS)*, pp. 2787–2795.

Bordes, A., J. Weston, R. Collobert, Y. Bengio, et al. (2011). Learning structured embeddings of knowledge bases. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, pp. 301–306.

Botha, J. A. and P. Blunsom (2014). Compositional morphology for word representations and language modelling. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Bottou, L. (1998). Online learning and stochastic approximations. *On-line learning in neural networks*, 9–42.

Bottou, L. (2012). Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade*, pp. 421–436. Springer.

Bottou, L., F. E. Curtis, and J. Nocedal (2016). Optimization methods for large-scale machine learning. *arXiv preprint arXiv:1606.04838*.

Bowman, S. R., L. Vilnis, O. Vinyals, A. Dai, R. Jozefowicz, and S. Bengio (2016). Generating sentences from a continuous space. In *Proceedings of the Conference on Natural Language Learning (CoNLL)*, pp. 10–21.

Boyd, S. and L. Vandenberghe (2004). *Convex Optimization*. New York: Cambridge University Press.

Branavan, S., H. Chen, J. Eisenstein, and R. Barzilay (2009). Learning document-level semantic properties from free-text annotations. *Journal of Artificial Intelligence Research 34*(2), 569–603.

Branavan, S. R., H. Chen, L. S. Zettlemoyer, and R. Barzilay (2009). Reinforcement learning for mapping instructions to actions. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 82–90.

Brown, P. F., P. V. Desouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai (1992). Class-based n-gram models of natural language. *Computational linguistics 18*(4), 467–479.

Brun, C. and C. Roux (2014). Décomposition des "hash tags" pour l'amélioration de la classification en polarité des "tweets". *Proceedings of Traitement Automatique des Langues Naturelles*, 473–478.

Bruni, E., N.-K. Tran, and M. Baroni (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research 49*(2014), 1–47.

Bullinaria, J. A. and J. P. Levy (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior research methods 39*(3), 510–526.

Bunescu, R. C. and R. J. Mooney (2005). A shortest path dependency kernel for relation extraction. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 724–731.

Bunescu, R. C. and M. Pasca (2006). Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 9–16.

Cai, Q. and A. Yates (2013). Large-scale semantic parsing via schema matching and lexicon extension. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 423–433.

Canny, J. (1987). A computational approach to edge detection. In *Readings in Computer Vision*, pp. 184–203. Elsevier.

Cappé, O. and E. Moulines (2009). On-line expectation–maximization algorithm for latent data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71*(3), 593–613.

Cardie, C. and K. Wagstaff (1999). Noun phrase coreference as clustering. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 82–89.

Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics 22*(2), 249–254.

Carpenter, B. (1997). *Type-logical semantics*. Cambridge, MA: MIT Press.

Carreras, X., M. Collins, and T. Koo (2008). Tag, dynamic programming, and the perceptron for efficient, feature-rich parsing. In *Proceedings of the Conference on Natural Language Learning (CoNLL)*, pp. 9–16.

Carreras, X. and L. Màrquez (2005). Introduction to the conll-2005 shared task: Semantic role labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, pp. 152–164. Association for Computational Linguistics.

Carroll, L. (1917). *Through the looking glass: And what Alice found there*. Chicago: Rand, McNally.

Chambers, N. and D. Jurafsky (2008). Jointly combining implicit constraints improves temporal ordering. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 698–706.

Chang, K.-W., A. Krishnamurthy, A. Agarwal, H. Daume III, and J. Langford (2015). Learning to search better than your teacher. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Chang, M.-W., L. Ratinov, and D. Roth (2007). Guiding semi-supervision with constraint-driven learning. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 280–287.

Chang, M.-W., L.-A. Ratinov, N. Rizzolo, and D. Roth (2008). Learning and inference with constraints. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, pp. 1513–1518.

Chapman, W. W., W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics 34*(5), 301–310.

Charniak, E. (1997). Statistical techniques for natural language parsing. *AI magazine 18*(4), 33–43.

Charniak, E. and M. Johnson (2005). Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 173–180.

Chelba, C. and A. Acero (2006). Adaptation of maximum entropy capitalizer: Little data can help a lot. *Computer Speech & Language 20*(4), 382–399.

Chen, D., J. Bolton, and C. D. Manning (2016). A thorough examination of the CNN/Daily Mail reading comprehension task. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Chen, D. and C. D. Manning (2014). A fast and accurate dependency parser using neural networks. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 740–750.

Chen, M., Z. Xu, K. Weinberger, and F. Sha (2012). Marginalized denoising autoencoders for domain adaptation. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Chen, S. F. and J. Goodman (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech & Language 13*(4), 359–393.

Chen, T. and C. Guestrin (2016). Xgboost: A scalable tree boosting system. In *Proceedings of Knowledge Discovery and Data Mining (KDD)*, pp. 785–794.

Chen, X., X. Qiu, C. Zhu, P. Liu, and X. Huang (2015). Long short-term memory neural networks for chinese word segmentation. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 1197–1206.

Chen, Z. and H. Ji (2009). Graph-based event coreference resolution. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing*, pp. 54–57. Association for Computational Linguistics.

Cheng, X. and D. Roth (2013). Relational inference for wikification. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 1787–1796.

Chiang, D., J. Graehl, K. Knight, A. Pauls, and S. Ravi (2010). Bayesian inference for finite-state transducers. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 447–455.

Cho, K., B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.

Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton & Co.

Choromanska, A., M. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun (2015). The loss surfaces of multilayer networks. In *Proceedings of Artificial Intelligence and Statistics (AISTATS)*, pp. 192–204.

Christensen, J., S. Soderland, O. Etzioni, et al. (2010). Semantic role labeling for open information extraction. In *Proceedings of the Workshop on Formalisms and Methodology for Learning by Reading*, pp. 52–60. Association for Computational Linguistics.

Chu, Y.-J. and T.-H. Liu (1965). On shortest arborescence of a directed graph. *Scientia Sinica 14*(10), 1396–1400.

Chung, C. and J. W. Pennebaker (2007). The psychological functions of function words. In K. Fiedler (Ed.), *Social communication*, pp. 343–359. New York and Hove: Psychology Press.

Church, K. W. (2000). Empirical estimates of adaptation: the chance of two Noriegas is closer to $p/2$ than $p^2$. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pp. 180–186.

Church, K. W. and P. Hanks (1990). Word association norms, mutual information, and lexicography. *Computational linguistics 16*(1), 22–29.

Ciaramita, M. and M. Johnson (2003). Supersense tagging of unknown nouns in wordnet. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 168–175.

Clark, K. and C. D. Manning (2015). Entity-centric coreference resolution with model stacking. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 1405–1415.

Clark, K. and C. D. Manning (2016). Improving coreference resolution by learning entity-level distributed representations. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Clark, P. (2015). Elementary school science and math tests as a driver for ai: take the aristo challenge! In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, pp. 4019–4021.

Clarke, J., D. Goldwasser, M.-W. Chang, and D. Roth (2010). Driving semantic parsing from the world's response. In *Proceedings of the Conference on Natural Language Learning (CoNLL)*, pp. 18–27.

Clarke, J. and M. Lapata (2008). Global inference for sentence compression: An integer linear programming approach. *Journal of Artificial Intelligence Research 31*, 399–429.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement 20*(1), 37–46.

Cohen, S. (2016). *Bayesian analysis in natural language processing*. Synthesis Lectures on Human Language Technologies. San Rafael, CA: Morgan & Claypool Publishers.

Cohen, S. B., C. Gómez-Rodríguez, and G. Satta (2012). Elimination of spurious ambiguity in transition-based dependency parsing. *CoRR abs/1206.6735*.

Collier, N., C. Nobata, and J.-i. Tsujii (2000). Extracting the names of genes and gene products with a hidden markov model. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pp. 201–207.

Collins, M. (1997). Three generative, lexicalised models for statistical parsing. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 16–23.

Collins, M. (2002). Discriminative training methods for hidden markov models: theory and experiments with perceptron algorithms. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 1–8.

Collins, M. (2013). Notes on natural language processing. `http://www.cs.columbia.edu/~mcollins/notes-spring2013.html`.

Collins, M. and T. Koo (2005). Discriminative reranking for natural language parsing. *Computational Linguistics 31*(1), 25–70.

Collins, M. and B. Roark (2004). Incremental parsing with the perceptron algorithm. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pp. 111. Association for Computational Linguistics.

Collobert, R., K. Kavukcuoglu, and C. Farabet (2011). Torch7: A matlab-like environment for machine learning. Technical Report EPFL-CONF-192376, EPFL.

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.

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.

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.

Cormen, T. H., C. E. Leiserson, R. L. Rivest, and C. Stein (2009). *Introduction to algorithms* (third ed.). MIT press.

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.

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.

Covington, M. A. (2001). A fundamental algorithm for dependency parsing. In *Proceedings of the 39th annual ACM southeast conference*, pp. 95–102.

Crammer, K. and Y. Singer (2001). Pranking with ranking. In *Neural Information Processing Systems (NIPS)*, pp. 641–647.

Creutz, M. and K. Lagus (2002). Unsupervised discovery of morphemes. In *Proceedings of the ACL-02 workshop on Morphological and phonological learning-Volume 6*, pp. 21–30. Association for Computational Linguistics.

Cross, J. and L. Huang (2016). Span-based constituency parsing with a structure-label system and provably optimal dynamic oracles. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 1–11.

Cucerzan, S. (2007). Large-scale named entity disambiguation based on wikipedia data. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.

Cui, H., R. Sun, K. Li, M.-Y. Kan, and T.-S. Chua (2005). Question answering passage retrieval using dependency relations. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 400–407. ACM.

Cui, Y., Z. Chen, S. Wei, S. Wang, T. Liu, and G. Hu (2017). Attention-over-attention neural networks for reading comprehension. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Culotta, A. and J. Sorensen (2004). Dependency tree kernels for relation extraction. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Culotta, A., M. Wick, and A. McCallum (2007). First-order probabilistic models for coreference resolution. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 81–88.

Curry, H. B. and R. Feys (1958). *Combinatory Logic*, Volume I. Amsterdam: North Holland.

Danescu-Niculescu-Mizil, C., M. Sudhof, D. Jurafsky, J. Leskovec, and C. Potts (2013). A computational approach to politeness with application to social factors. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 250–259.

Das, D., D. Chen, A. F. Martins, N. Schneider, and N. A. Smith (2014). Frame-semantic parsing. *Computational Linguistics 40*(1), 9–56.

Daumé III, H. (2007). Frustratingly easy domain adaptation. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Daumé III, H., J. Langford, and D. Marcu (2009). Search-based structured prediction. *Machine learning 75*(3), 297–325.

Daumé III, H. and D. Marcu (2005). A large-scale exploration of effective global features for a joint entity detection and tracking model. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 97–104.

Dauphin, Y. N., R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio (2014). Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Neural Information Processing Systems (NIPS)*, pp. 2933–2941.

Davidson, D. (1967). The logical form of action sentences. In N. Rescher (Ed.), *The Logic of Decision and Action*. Pittsburgh: University of Pittsburgh Press.

De Marneffe, M.-C. and C. D. Manning (2008). The stanford typed dependencies representation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pp. 1–8. Association for Computational Linguistics.

Dean, J. and S. Ghemawat (2008). Mapreduce: simplified data processing on large clusters. *Communications of the ACM 51*(1), 107–113.

Deerwester, S. C., S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman (1990). Indexing by latent semantic analysis. *JASIS 41*(6), 391–407.

Dehdari, J. (2014). *A Neurophysiologically-Inspired Statistical Language Model*. Ph. D. thesis, The Ohio State University.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–38.

Denis, P. and J. Baldridge (2007). A ranking approach to pronoun resolution. In *IJCAI*.

Denis, P. and J. Baldridge (2008). Specialized models and ranking for coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, Stroudsburg, PA, USA, pp. 660–669. Association for Computational Linguistics.

Denis, P. and J. Baldridge (2009). Global joint models for coreference resolution and named entity classification. *Procesamiento del Lenguaje Natural 42*.

Dhingra, B., H. Liu, Z. Yang, W. W. Cohen, and R. Salakhutdinov (2017). Gated-attention readers for text comprehension. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Diaconis, P. and B. Skyrms (2017). *Ten Great Ideas About Chance*. Princeton University Press.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation 10*(7), 1895–1923.

Dietterich, T. G., R. H. Lathrop, and T. Lozano-Pérez (1997). Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence 89*(1), 31–71.

Dimitrova, L., N. Ide, V. Petkevic, T. Erjavec, H. J. Kaalep, and D. Tufis (1998). Multext-east: Parallel and comparable corpora and lexicons for six central and eastern european languages. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pp. 315–319. Association for Computational Linguistics.

Doddington, G. R., A. Mitchell, M. A. Przybocki, L. A. Ramshaw, S. Strassel, and R. M. Weischedel (2004). The automatic content extraction (ace) program-tasks, data, and evaluation. In *Proceedings of the Language Resources and Evaluation Conference*, pp. 837–840.

dos Santos, C., B. Xiang, and B. Zhou (2015). Classifying relations by ranking with convolutional neural networks. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 626–634.

Dowty, D. (1991). Thematic proto-roles and argument selection. *Language*, 547–619.

Dredze, M., P. McNamee, D. Rao, A. Gerber, and T. Finin (2010). Entity disambiguation for knowledge base population. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pp. 277–285. Association for Computational Linguistics.

Dredze, M., M. J. Paul, S. Bergsma, and H. Tran (2013). Carmen: A Twitter geolocation system with applications to public health. In *AAAI workshop on expanding the boundaries of health informatics using AI (HIAI)*, pp. 20–24.

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.

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*.

Durrett, G. and D. Klein (2013). Easy victories and uphill battles in coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Durrett, G. and D. Klein (2015). Neural crf parsing. In *Proceedings of the Association for Computational Linguistics (ACL)*.

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.

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.

Edmonds, J. (1967). Optimum branchings. *Journal of Research of the National Bureau of Standards B 71*(4), 233–240.

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*.

Eisner, J. (2000). Bilexical grammars and their cubic-time parsing algorithms. In *Advances in probabilistic and other parsing technologies*, pp. 29–61. Springer.

Eisner, J. (2002). Parameter estimation for probabilistic finite-state transducers. In *acl*, pp. 1–8.

Eisner, J. M. (1996). Three new probabilistic models for dependency parsing: An exploration. In *COLING*, pp. 340–345.

Ekman, P. (1992). Are there basic emotions? *Psychological Review 99*(3), 550–553.

Esuli, A. and F. Sebastiani (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In *LREC*, Volume 6, pp. 417–422. Citeseer.

Etzioni, O., A. Fader, J. Christensen, S. Soderland, and M. Mausam (2011). Open information extraction: The second generation. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 3–10.

Faruqui, M., J. Dodge, S. K. Jauhar, C. Dyer, E. Hovy, and N. A. Smith (2015). Retrofitting word vectors to semantic lexicons. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Faruqui, M. and C. Dyer (2014). Improving vector space word representations using multilingual correlation. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 462–471.

Faruqui, M., R. McDonald, and R. Soricut (2016). Morpho-syntactic lexicon generation using graph-based semi-supervised learning. *Transactions of the Association for Computational Linguistics 4*, 1–16.

Faruqui, M., Y. Tsvetkov, P. Rastogi, and C. Dyer (2016, August). Problems with evaluation of word embeddings using word similarity tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, Berlin, Germany, pp. 30–35. Association for Computational Linguistics.

Fellbaum, C. (2010). *WordNet*. Springer.

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.

Fernandes, E. R., C. N. dos Santos, and R. L. Milidiú (2014). Latent trees for coreference resolution. *Computational Linguistics*.

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.

Figueiredo, M., J. Graça, A. Martins, M. Almeida, and L. P. Coelho (2013). LXMLS lab guide. `http://lxmls.it.pt/2013/guide.pdf`.

Fillmore, C. J. (1968). The case for case. In E. Bach and R. Harms (Eds.), *Universals in linguistic theory*. Holt, Rinehart, and Winston.

Fillmore, C. J. (1976). Frame semantics and the nature of language. *Annals of the New York Academy of Sciences 280*(1), 20–32.

Fillmore, C. J. and C. Baker (2009). A frames approach to semantic analysis. In *The Oxford Handbook of Linguistic Analysis*. Oxford University Press.

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.

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.

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.

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.

Finkel, J. R. and C. D. Manning (2008). Enforcing transitivity in coreference resolution. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pp. 45–48. Association for Computational Linguistics.

Finkelstein, L., E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin (2002). Placing search in context: The concept revisited. *ACM Transactions on Information Systems 20*(1), 116–131.

Firth, J. R. (1957). *Papers in Linguistics 1934-1951*. Oxford University Press.

Flanigan, J., S. Thomson, J. Carbonell, C. Dyer, and N. A. Smith (2014). A discriminative graph-based parser for the abstract meaning representation. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 1426–1436.

Francis, W. N. (1964). A standard sample of present-day English for use with digial computers. Report to the U.S Office of Education on Cooperative Research Project No. E-007.

Freund, Y. and R. E. Schapire (1999). Large margin classification using the perceptron algorithm. *Machine learning 37*(3), 277–296.

Fromkin, V., R. Rodman, and N. Hyams (2013). *An introduction to language*. Cengage Learning.

Fundel, K., R. Küffner, and R. Zimmer (2007). Relex – relation extraction using dependency parse trees. *Bioinformatics 23*(3), 365–371.

Gabrilovich, E. and S. Markovitch (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, Volume 7, pp. 1606–1611.

Gage, P. (1994). A new algorithm for data compression. *The C Users Journal 12*(2), 23–38.

Gale, W. A., K. W. Church, and D. Yarowsky (1992). One sense per discourse. In *Proceedings of the workshop on Speech and Natural Language*, pp. 233–237. Association for Computational Linguistics.

Ganchev, K. and M. Dredze (2008). Small statistical models by random feature mixing. In *Proceedings of the ACL08 HLT Workshop on Mobile Language Processing*, pp. 19–20.

Ganchev, K., J. Graça, J. Gillenwater, and B. Taskar (2010). Posterior regularization for structured latent variable models. *The Journal of Machine Learning Research 11*, 2001–2049.

Ganin, Y., E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky (2016). Domain-adversarial training of neural networks. *Journal of Machine Learning Research 17*(59), 1–35.

Gao, J., G. Andrew, M. Johnson, and K. Toutanova (2007). A comparative study of parameter estimation methods for statistical natural language processing. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 824–831.

Ge, D., X. Jiang, and Y. Ye (2011). A note on the complexity of l p minimization. *Mathematical programming 129*(2), 285–299.

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.

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)*.

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.

Geach, P. T. (1962). *Reference and generality: An examination of some medieval and modern theories*. Cornell University Press.

Gildea, D. and D. Jurafsky (2002). Automatic labeling of semantic roles. *Computational linguistics 28*(3), 245–288.

Gimpel, K., N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, 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.

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.

Glorot, X., A. Bordes, and Y. Bengio (2011). Deep sparse rectifier networks. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. JMLR W&CP Volume*, Volume 15, pp. 315–323.

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.

Goldberg, Y. (2017a, June). An adversarial review of "adversarial generation of natural language". https://medium.com/@yoav.goldberg/an-adversarial-review-of-adversarial-generation-of-natural-language-409ac3378bd7.

Goldberg, Y. (2017b). *Neural Network Methods for Natural Language Processing*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Goldberg, Y. and M. Elhadad (2010). An efficient algorithm for easy-first non-directional dependency parsing. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 742–750.

Goldberg, Y. and J. Nivre (2012). A dynamic oracle for arc-eager dependency parsing. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pp. 959–976.

Goldberg, Y., K. Zhao, and L. Huang (2013). Efficient implementation of beam-search incremental parsers. In *ACL (2)*, pp. 628–633.

Goldwater, S. and T. Griffiths (2007). A fully bayesian approach to unsupervised part-of-speech tagging. In *Annual meeting-association for computational linguistics*, Volume 45.

Goodfellow, I., Y. Bengio, and A. Courville (2016). *Deep learning*. MIT Press.

Goodman, J. T. (2001). A bit of progress in language modeling. *Computer Speech & Language 15*(4), 403–434.

Gouws, S., D. Metzler, C. Cai, and E. Hovy (2011). Contextual bearing on linguistic variation in social media. In *LASM*.

Graves, A. and J. Schmidhuber (2005). Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks 18*(5), 602–610.

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.

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.

Grishman, R. (2015). Information extraction. *IEEE Intelligent Systems 30*(5), 8–15.

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.

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.

Groenendijk, J. and M. Stokhof (1991). Dynamic predicate logic. *Linguistics and philosophy 14*(1), 39–100.

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.

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.

Haghighi, A. and D. Klein (2007). Unsupervised coreference resolution in a nonparametric bayesian model. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Haghighi, 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.

Haghighi, A. and D. Klein (2010). Coreference resolution in a modular, entity-centered model. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 385–393.

Hajič, J. and B. Hladká (1998). Tagging inflective languages: Prediction of morphological categories for a rich, structured tagset. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 483–490.

Hammerton, J. (2003). Named entity recognition with long short-term memory. In *Proceedings of the Conference on Natural Language Learning (CoNLL)*, pp. 172–175.

Han, X. and L. Sun (2012). An entity-topic model for entity linking. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 105–115.

Han, X., L. Sun, and J. Zhao (2011). Collective entity linking in web text: a graph-based method. In *Proceedings of ACM SIGIR conference on Research and development in information retrieval*, pp. 765–774.

Hannak, A., E. Anderson, L. F. Barrett, S. Lehmann, A. Mislove, and M. Riedewald (2012). Tweetin'in the rain: Exploring societal-scale effects of weather on mood. In *Proceedings of the International Conference on Web and Social Media (ICWSM)*.

Haspelmath, M. and A. Sims (2013). *Understanding morphology*. Routledge.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The elements of statistical learning* (Second ed.). New York: Springer.

Hatzivassiloglou, V. and K. R. McKeown (1997). Predicting the semantic orientation of adjectives. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 174–181.

Hayes, A. F. and K. Krippendorff (2007). Answering the call for a standard reliability measure for coding data. *Communication methods and measures 1*(1), 77–89.

He, K., X. Zhang, S. Ren, and J. Sun (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pp. 1026–1034.

He, K., X. Zhang, S. Ren, and J. Sun (2016). Deep residual learning for image recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pp. 770–778.

He, L., K. Lee, M. Lewis, and L. Zettlemoyer (2017). Deep semantic role labeling: What works and what's next. In *Proceedings of the Association for Computational Linguistics (ACL)*.

He, Z., S. Liu, M. Li, M. Zhou, L. Zhang, and H. Wang (2013). Learning entity representation for entity disambiguation. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 30–34.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pp. 539–545. Association for Computational Linguistics.

Hendrickx, I., S. N. Kim, Z. Kozareva, P. Nakov, D. Ó Séaghdha, S. Padó, M. Pennacchiotti, L. Romano, and S. Szpakowicz (2009). Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pp. 94–99. Association for Computational Linguistics.

Hermann, K. M., T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom (2015). Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pp. 1693–1701.

Hill, F., A. Bordes, S. Chopra, and J. Weston (2016). The goldilocks principle: Reading children's books with explicit memory representations. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Hindle, D. and M. Rooth (1993). Structural ambiguity and lexical relations. *Computational linguistics 19*(1), 103–120.

Hirschman, L. and R. Gaizauskas (2001). Natural language question answering: the view from here. *natural language engineering 7*(4), 275–300.

Hirschman, L., M. Light, E. Breck, and J. D. Burger (1999). Deep read: A reading comprehension system. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 325–332.

Hobbs, J. R. (1978). Resolving pronoun references. *Lingua 44*(4), 311–338.

Hobbs, J. R., D. Appelt, J. Bear, D. Israel, M. Kameyama, M. Stickel, and M. Tyson (1997). Fastus: A cascaded finite-state transducer for extracting information from natural-language text. *Finite-state language processing*, 383–406.

Hochreiter, S. and J. Schmidhuber (1997). Long short-term memory. *Neural computation 9*(8), 1735–1780.

Hockenmaier, J. and M. Steedman (2007). Ccgbank: a corpus of ccg derivations and dependency structures extracted from the penn treebank. *Computational Linguistics 33*(3), 355–396.

Hoffart, J., M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum (2011). Robust disambiguation of named entities in text. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 782–792.

Hoffmann, R., C. Zhang, X. Ling, L. Zettlemoyer, and D. S. Weld (2011). Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 541–550.

Holmstrom, L. and P. Koistinen (1992). Using additive noise in back-propagation training. *IEEE Transactions on Neural Networks 3*(1), 24–38.

Hovy, E. and J. Lavid (2010). Towards a 'science' of corpus annotation: a new methodological challenge for corpus linguistics. *International journal of translation 22*(1), 13–36.

Hsu, D., S. M. Kakade, and T. Zhang (2012). A spectral algorithm for learning hidden markov models. *Journal of Computer and System Sciences 78*(5), 1460–1480.

Hu, M. and B. Liu (2004). Mining and summarizing customer reviews. In *Proceedings of Knowledge Discovery and Data Mining (KDD)*, pp. 168–177.

Huang, F. and A. Yates (2012). Biased representation learning for domain adaptation. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 1313–1323.

Huang, L., S. Fayong, and Y. Guo (2012). Structured perceptron with inexact search. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 142–151.

Huang, Y. (2015). *Pragmatics* (Second ed.). Oxford Textbooks in Linguistics. Oxford University Press.

Huang, Z., W. Xu, and K. Yu (2015). Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Huffman, D. A. (1952). A method for the construction of minimum-redundancy codes. *Proceedings of the IRE 40*(9), 1098–1101.

Humphreys, K., R. Gaizauskas, and S. Azzam (1997). Event coreference for information extraction. In *Proceedings of a Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, pp. 75–81. Association for Computational Linguistics.

Ide, N. and Y. Wilks (2006). Making sense about sense. In *Word sense disambiguation*, pp. 47–73. Springer.

Ioffe, S. and C. Szegedy (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 448–456.

Iyyer, M., V. Manjunatha, J. Boyd-Graber, and H. Daumé III (2015). Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 1681–1691.

Jeong, M., C.-Y. Lin, and G. G. Lee (2009). Semi-supervised speech act recognition in emails and forums. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 1250–1259.

Ji, H. and R. Grishman (2011). Knowledge base population: Successful approaches and challenges. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 1148–1158.

Ji, Y. and J. Eisenstein (2015, June). One vector is not enough: Entity-augmented distributional semantics for discourse relations. *Transactions of the Association for Computational Linguistics (TACL)*.

Ji, Y., G. Haffari, and J. Eisenstein (2016). A latent variable recurrent neural network for discourse relation language models. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Ji, Y., C. Tan, S. Martschat, Y. Choi, and N. A. Smith (2017). Dynamic entity representations in neural language models. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 1831–1840.

Jiang, L., M. Yu, M. Zhou, X. Liu, and T. Zhao (2011). Target-dependent twitter sentiment classification. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 151–160.

Jing, H. (2000). Sentence reduction for automatic text summarization. In *Proceedings of the sixth conference on Applied natural language processing*, pp. 310–315. Association for Computational Linguistics.

Joachims, T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of Knowledge Discovery and Data Mining (KDD)*, pp. 133–142.

Jockers, M. L. (2015). Szuzhet? `http:bla.bla.com`.

Johnson, M. (1998). Pcfg models of linguistic tree representations. *Computational Linguistics 24*(4), 613–632.

Johnson, R. and T. Zhang (2017). Deep pyramid convolutional neural networks for text categorization. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 562–570.

Joshi, A. K. (1985). How much context-sensitivity is required to provide reasonable structural descriptions? – tree adjoining grammars. In *Natural Language Processing – Theoretical, Computational and Psychological Perspective*. New York, NY: Cambridge University Press.

Joshi, A. K. and Y. Schabes (1997). Tree-adjoining grammars. In *Handbook of formal languages*, pp. 69–123. Springer.

Joshi, A. K., K. V. Shanker, and D. Weir (1991). The convergence of mildly context-sensitive grammar formalisms. In *Foundational Issues in Natural Language Processing*. Cambridge MA: MIT Press.

Jozefowicz, R., W. Zaremba, and I. Sutskever (2015). An empirical exploration of recurrent network architectures. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 2342–2350.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science 20*(2), 137–194.

Jurafsky, D. and J. H. Martin (2009). *Speech and Language Processing* (Second ed.). Prentice Hall.

Jurafsky, D. and J. H. Martin (2018). *Speech and Language Processing* (Third ed.). Prentice Hall.

Kadlec, R., M. Schmid, O. Bajgar, and J. Kleindienst (2016). Text understanding with the attention sum reader network. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 908–918.

Kalchbrenner, N. and P. Blunsom (2013, August). Recurrent convolutional neural networks for discourse compositionality. In *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*, Sofia, Bulgaria, pp. 119–126. Association for Computational Linguistics.

Kalchbrenner, N., E. Grefenstette, and P. Blunsom (2014). A convolutional neural network for modelling sentences. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 655–665.

Karlsson, F. (2007). Constraints on multiple center-embedding of clauses. *Journal of Linguistics 43*(02), 365–392.

Kate, R. J., Y. W. Wong, and R. J. Mooney (2005). Learning to transform natural to formal languages. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*.

Kehler, A. (2007). Rethinking the SMASH approach to pronoun interpretation. In *Interdisciplinary perspectives on reference processing*, New Directions in Cognitive Science Series, pp. 95–122. Oxford University Press.

Kilgarriff, A. (1997). I don't believe in word senses. *Computers and the Humanities 31*(2), 91–113.

Kim, M.-J. (2002). Does korean have adjectives? *MIT Working Papers in Linguistics 43*, 71–89.

Kim, S.-M. and E. Hovy (2006, July). Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, Sydney, Australia, pp. 1–8. Association for Computational Linguistics.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 1746–1751.

Kim, Y., Y. Jernite, D. Sontag, and A. M. Rush (2016). Character-aware neural language models. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*.

Kingma, D. and J. Ba (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P. and M. Welling (2014). Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Kiperwasser, E. and Y. Goldberg (2016). Simple and accurate dependency parsing using bidirectional lstm feature representations. *Transactions of the Association for Computational Linguistics 4*, 313–327.

Kipper-Schuler, K. (2005). *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph. D. thesis, Computer and Information Science, University of Pennsylvania.

Kiros, R., R. Salakhutdinov, and R. Zemel (2014). Multimodal neural language models. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 595–603.

Kiros, R., Y. Zhu, R. Salakhudinov, R. S. Zemel, A. Torralba, R. Urtasun, and S. Fidler (2015). Skip-thought vectors. In *Neural Information Processing Systems (NIPS)*.

Klein, D. and C. D. Manning (2003). Accurate unlexicalized parsing. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 423–430.

Klein, D. and C. D. Manning (2004). Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Klementiev, A., I. Titov, and B. Bhattarai (2012). Inducing crosslingual distributed representations of words. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pp. 1459–1474.

Klenner, M. (2007). Enforcing consistency on coreference sets. In *Recent Advances in Natural Language Processing (RANLP)*, pp. 323–328.

Knight, K. and J. May (2009). Applications of weighted automata in natural language processing. In *Handbook of Weighted Automata*, pp. 571–596. Springer.

Koo, T., X. Carreras, and M. Collins (2008, jun). Simple semi-supervised dependency parsing. In *Proceedings of ACL-08: HLT*, Columbus, Ohio, pp. 595–603. Association for Computational Linguistics.

Koo, T. and M. Collins (2005). Hidden-variable models for discriminative reranking. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 507–514.

Koo, T. and M. Collins (2010). Efficient third-order dependency parsers. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Koo, T., A. Globerson, X. Carreras, and M. Collins (2007). Structured prediction models via the matrix-tree theorem. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 141–150.

Kovach, B. and T. Rosenstiel (2014). *The elements of journalism: What newspeople should know and the public should expect*. Three Rivers Press.

Krishnamurthy, J. (2016). Probabilistic models for learning a semantic parser lexicon. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 606–616.

Krishnamurthy, J. and T. M. Mitchell (2012). Weakly supervised training of semantic parsers. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 754–765.

Kübler, S., R. McDonald, and J. Nivre (2009). Dependency parsing. *Synthesis Lectures on Human Language Technologies 1*(1), 1–127.

Kuhlmann, M. and J. Nivre (2010). Transition-based techniques for non-projective dependency parsing. *Northern European Journal of Language Technology (NEJLT) 2*(1), 1–19.

Kummerfeld, J. K., T. Berg-Kirkpatrick, and D. Klein (2015). An empirical analysis of optimization for max-margin NLP. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.

Kwiatkowski, T., S. Goldwater, L. Zettlemoyer, and M. Steedman (2012). A probabilistic model of syntactic and semantic acquisition from child-directed utterances and their meanings. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 234–244.

Lafferty, J., A. McCallum, and F. Pereira (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *icml*.

Lakoff, G. (1973). Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Journal of philosophical logic 2*(4), 458–508.

Lample, G., M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer (2016). Neural architectures for named entity recognition. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 260–270.

Lappin, S. and H. J. Leass (1994). An algorithm for pronominal anaphora resolution. *Computational linguistics 20*(4), 535–561.

Lari, K. and S. J. Young (1990). The estimation of stochastic context-free grammars using the inside-outside algorithm. *Computer speech & language 4*(1), 35–56.

Law, E. and L. v. Ahn (2011). Human computation. *Synthesis Lectures on Artificial Intelligence and Machine Learning 5*(3), 1–121.

LeCun, Y. and Y. Bengio (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks 3361*.

LeCun, Y., L. Bottou, G. B. Orr, and K.-R. Müller (1998). Efficient backprop. In *Neural networks: Tricks of the trade*, pp. 9–50. Springer.

Lee, C. M. and S. S. Narayanan (2005). Toward detecting emotions in spoken dialogs. *IEEE transactions on speech and audio processing 13*(2), 293–303.

Lee, H., A. Chang, Y. Peirsman, N. Chambers, M. Surdeanu, and D. Jurafsky (2013). Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics 39*(4), 885–916.

Lee, H., Y. Peirsman, A. Chang, N. Chambers, M. Surdeanu, and D. Jurafsky (2011). Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. In *Proceedings of the Conference on Natural Language Learning (CoNLL)*, pp. 28–34. Association for Computational Linguistics.

Lee, K., L. He, M. Lewis, and L. Zettlemoyer (2017). End-to-end neural coreference resolution. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.

Lenat, D. B., R. V. Guha, K. Pittman, D. Pratt, and M. Shepherd (1990). Cyc: toward programs with common sense. *Communications of the ACM 33*(8), 30–49.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: 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.

Levy, O. and Y. Goldberg (2014). Dependency-based word embeddings. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 302–308.

Levy, O., Y. Goldberg, and I. Dagan (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics 3*, 211–225.

Lewis, M. and M. Steedman (2013). Combined distributional and logical semantics. *Transactions of the Association for Computational Linguistics 1*, 179–192.

Li, J. and D. Jurafsky (2015). Do multi-sense embeddings improve natural language understanding? In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 1722–1732.

Li, J., M.-T. Luong, and D. Jurafsky (2015). A hierarchical neural autoencoder for paragraphs and documents. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.

Li, J., T. Luong, D. Jurafsky, and E. Hovy (2015). When are tree structures necessary for deep learning of representations? In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 2304–2314.

Li, Q., S. Anzaroot, W.-P. Lin, X. Li, and H. Ji (2011). Joint inference for cross-document information extraction. In *Proceedings of the International Conference on Information and Knowledge Management (CIKM)*, pp. 2225–2228.

Li, Q., H. Ji, and L. Huang (2013). Joint event extraction via structured prediction with global features. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 73–82.

Liang, P., A. Bouchard-Côté, D. Klein, and B. Taskar (2006). An end-to-end discriminative approach to machine translation. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 761–768.

Liang, P., M. I. Jordan, and D. Klein (2013). Learning dependency-based compositional semantics. *Computational Linguistics 39*(2), 389–446.

Liang, P. and D. Klein (2009). Online em for unsupervised models. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 611–619.

Liang, P., S. Petrov, M. I. Jordan, and D. Klein (2007). The infinite pcfg using hierarchical dirichlet processes. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 688–697.

Liang, P. and C. Potts (2015). Bringing machine learning and compositional semantics together. *Annual Review of Linguistics 1*(1), 355–376.

Lieber, R. (2015). *Introducing morphology*. Cambridge University Press.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics-Volume 2*, pp. 768–774. Association for Computational Linguistics.

Ling, W., C. Dyer, A. Black, and I. Trancoso (2015). Two/too simple adaptations of word2vec for syntax problems. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Ling, W., T. Luís, L. Marujo, R. F. Astudillo, S. Amir, C. Dyer, A. W. Black, and I. Trancoso (2015). Finding function in form: Compositional character models for open vocabulary word representation. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.

Ling, X., S. Singh, and D. S. Weld (2015). Design challenges for entity linking. *Transactions of the Association for Computational Linguistics 3*, 315–328.

Linguistic Data Consortium (2005, July). ACE (automatic content extraction) English annotation guidelines for relations. Technical Report Version 5.8.3, Linguistic Data Consortium.

Liu, B. (2015). *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press.

Liu, D. C. and J. Nocedal (1989). On the limited memory BFGS method for large scale optimization. *Mathematical programming 45*(1-3), 503–528.

Loveland, D. W. (2016). *Automated Theorem Proving: a logical basis*. Elsevier.

Luo, X. (2005). On coreference resolution performance metrics. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 25–32.

Luo, X., A. Ittycheriah, H. Jing, N. Kambhatla, and S. Roukos (2004). A mention-synchronous coreference resolution algorithm based on the bell tree. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Luong, M.-T., R. Socher, and C. D. Manning (2013). Better word representations with recursive neural networks for morphology. *CoNLL-2013 104*.

Maas, A. L., A. Y. Hannun, and A. Y. Ng (2013). Rectifier nonlinearities improve neural network acoustic models. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Mani, I., M. Verhagen, B. Wellner, C. M. Lee, and J. Pustejovsky (2006). Machine learning of temporal relations. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 753–760.

Manning, C. D. (2016). Computational linguistics and deep learning. *Computational Linguistics 41*(4).

Manning, C. D., P. Raghavan, H. Schütze, et al. (2008). *Introduction to information retrieval*, Volume 1. Cambridge university press.

Manning, C. D. and H. Schütze (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: MIT press.

Marcus, M. P., M. A. Marcinkiewicz, and B. Santorini (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics 19*(2), 313–330.

Maron, O. and T. Lozano-Pérez (1998). A framework for multiple-instance learning. In *Neural Information Processing Systems (NIPS)*, pp. 570–576.

Márquez, G. G. (1970). *One Hundred Years of Solitude*. Harper & Row. English translation by Gregory Rabassa.

Martins, A. F. T., N. A. Smith, and E. P. Xing (2009). Concise integer linear programming formulations for dependency parsing. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 342–350.

Martins, A. F. T., N. A. Smith, E. P. Xing, P. M. Q. Aguiar, and M. A. T. Figueiredo (2010). Turbo parsers: Dependency parsing by approximate variational inference. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 34–44.

Matsuzaki, T., Y. Miyao, and J. Tsujii (2005). Probabilistic cfg with latent annotations. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 75–82.

Matthiessen, C. and J. A. Bateman (1991). *Text generation and systemic-functional linguistics: experiences from English and Japanese*. Pinter Publishers.

May, J. (2016). Semeval-2016 task 8: Meaning representation parsing. In *Proceedings of SemEval*, pp. 1063–1073.

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.

McCallum, A. and B. Wellner (2004). Conditional models of identity uncertainty with application to noun coreference. In *NIPS*, pp. 905–912.

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.

McDonald, R., K. Hannan, T. Neylon, M. Wells, and J. Reynar (2007). Structured models for fine-to-coarse sentiment analysis. In *Proceedings of ACL*.

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)*.

McDonald, R., F. Pereira, K. Ribarov, and J. Hajič (2005). Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 523–530.

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.

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.

Miao, Y., L. Yu, and P. Blunsom (2016). Neural variational inference for text processing. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Mihalcea, R., T. A. Chklovski, and A. Kilgarriff (2004, July). The senseval-3 english lexical sample task. In *Proceedings of SENSEVAL-3*, Barcelona, Spain, pp. 25–28. Association for Computational Linguistics.

Mihalcea, R. and D. Radev (2011). *Graph-based natural language processing and information retrieval*. Cambridge University Press.

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*.

Mikolov, T., A. Deoras, D. Povey, L. Burget, and J. Cernocky (2011). Strategies for training large scale neural network language models. In *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*, pp. 196–201. IEEE.

Mikolov, T., M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur (2010). Recurrent neural network based language model. In *INTERSPEECH*, pp. 1045–1048.

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.

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.

Miller, M., C. Sathi, D. Wiesenthal, J. Leskovec, and C. Potts (2011). Sentiment flow through hyperlink networks. In *Proceedings of the International Conference on Web and Social Media (ICWSM)*.

Miller, S., J. Guinness, and A. Zamanian (2004). Name tagging with word clusters and discriminative training. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 337–342.

Milne, D. and I. H. Witten (2008). Learning to link with wikipedia. In *Proceedings of the International Conference on Information and Knowledge Management (CIKM)*, pp. 509–518. ACM.

Minka, T. P. (1999). From hidden markov models to linear dynamical systems. Tech. Rep. 531, Vision and Modeling Group of Media Lab, MIT.

Minsky, M. (1974). A framework for representing knowledge. Technical Report 306, MIT AI Laboratory.

Minsky, M. and S. Papert (1969). *Perceptrons*. MIT press.

Mintz, M., S. Bills, R. Snow, and D. Jurafsky (2009). Distant supervision for relation extraction without labeled data. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 1003–1011.

Mirza, P., R. Sprugnoli, S. Tonelli, and M. Speranza (2014). Annotating causality in the tempeval-3 corpus. In *Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL)*, pp. 10–19.

Misra, D. K. and Y. Artzi (2016). Neural shift-reduce ccg semantic parsing. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.

Mitchell, J. and M. Lapata (2010). Composition in distributional models of semantics. *Cognitive Science 34*(8), 1388–1429.

Miwa, M. and M. Bansal (2016). End-to-end relation extraction using lstms on sequences and tree structures. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 1105–1116.

Mnih, A. and G. Hinton (2007). Three new graphical models for statistical language modelling. In *Proceedings of the 24th international conference on Machine learning*, ICML '07, New York, NY, USA, pp. 641–648. ACM.

Mnih, A. and G. E. Hinton (2008). A scalable hierarchical distributed language model. In *Neural Information Processing Systems (NIPS)*, pp. 1081–1088.

Mnih, A. and Y. W. Teh (2012). A fast and simple algorithm for training neural probabilistic language models. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Mohammad, S. M. and P. D. Turney (2013). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence 29*(3), 436–465.

Mohri, M., F. Pereira, and M. Riley (2002). Weighted finite-state transducers in speech recognition. *Computer Speech & Language 16*(1), 69–88.

Mohri, M., A. Rostamizadeh, and A. Talwalkar (2012). *Foundations of machine learning*. MIT press.

Montague, R. (1973). The proper treatment of quantification in ordinary english. In *Approaches to natural language*, pp. 221–242. Springer.

Morante, R. and E. Blanco (2012). *sem 2012 shared task: Resolving the scope and focus of negation. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pp. 265–274. Association for Computational Linguistics.

Morante, R. and W. Daelemans (2009). Learning the scope of hedge cues in biomedical texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pp. 28–36. Association for Computational Linguistics.

Morante, R. and C. Sporleder (2012). Modality and negation: An introduction to the special issue. *Computational linguistics 38*(2), 223–260.

Mostafazadeh, N., A. Grealish, N. Chambers, J. Allen, and L. Vanderwende (2016, June). Caters: Causal and temporal relation scheme for semantic annotation of event structures. In *Proceedings of the Fourth Workshop on Events*, San Diego, California, pp. 51–61. Association for Computational Linguistics.

Mueller, T., H. Schmid, and H. Schütze (2013). Efficient higher-order CRFs for morphological tagging. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 322–332.

Müller, C. and M. Strube (2006). Multi-level annotation of linguistic data with mmax2. *Corpus technology and language pedagogy: New resources, new tools, new methods 3*, 197–214.

Muralidharan, A. and M. A. Hearst (2013). Supporting exploratory text analysis in literature study. *Literary and linguistic computing 28*(2), 283–295.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press.

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.

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR) 41*(2), 10.

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.

Nemirovski, A. and D. Yudin (1978). On Cezari's convergence of the steepest descent method for approximating saddle points of convex-concave functions. *Soviet Math. Dokl.*.

Neubig, G., C. Dyer, Y. Goldberg, A. Matthews, W. Ammar, A. Anastasopoulos, M. Ballesteros, D. Chiang, D. Clothiaux, T. Cohn, K. Duh, M. Faruqui, C. Gan, D. Garrette, Y. Ji, L. Kong, A. Kuncoro, G. Kumar, C. Malaviya, P. Michel, Y. Oda, M. Richardson, N. Saphra, S. Swayamdipta, and P. Yin (2017). Dynet: The dynamic neural network toolkit.

Neubig, G., Y. Goldberg, and C. Dyer (2017). On-the-fly operation batching in dynamic computation graphs. In *Neural Information Processing Systems (NIPS)*.

Neuhaus, P. and N. Bröker (1997). The complexity of recognition of linguistically adequate dependency grammars. In *eacl*, pp. 337–343.

Newman, D., C. Chemudugunta, and P. Smyth (2006). Statistical entity-topic models. In *Proceedings of Knowledge Discovery and Data Mining (KDD)*, pp. 680–686.

Ng, V. (2010). Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pp. 1396–1411. Association for Computational Linguistics.

Nguyen, D. and A. S. Dogruöz (2013). Word level language identification in online multilingual communication. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.

Nigam, K., A. K. McCallum, S. Thrun, and T. Mitchell (2000). Text classification from labeled and unlabeled documents using em. *Machine learning 39*(2-3), 103–134.

Nivre, J. (2008). Algorithms for deterministic incremental dependency parsing. *Computational Linguistics 34*(4), 513–553.

Nivre, J., M.-C. de Marneffe, F. Ginter, Y. Goldberg, J. Hajič, C. D. Manning, R. McDonald, S. Petrov, S. Pyysalo, N. Silveira, R. Tsarfaty, and D. Zeman (2016, may). Universal dependencies v1: A multilingual treebank collection. In N. C. C. Chair), K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Nivre, J. and J. Nilsson (2005). Pseudo-projective dependency parsing. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pp. 99–106. Association for Computational Linguistics.

Novikoff, A. B. (1962). On convergence proofs on perceptrons. In *Proceedings of the Symposium on the Mathematical Theory of Automata*, Volume 12, pp. 615—622.

O'Connor, B., M. Krieger, and D. Ahn (2010). Tweetmotif: Exploratory search and topic summarization for twitter. In *Proceedings of the International Conference on Web and Social Media (ICWSM)*, pp. 384–385.

Oflazer, K. and Ì. Kuruöz (1994). Tagging and morphological disambiguation of turkish text. In *Proceedings of the fourth conference on Applied natural language processing*, pp. 144–149. Association for Computational Linguistics.

Ohta, T., Y. Tateisi, and J.-D. Kim (2002). The genia corpus: An annotated research abstract corpus in molecular biology domain. In *Proceedings of the second international conference on Human Language Technology Research*, pp. 82–86. Morgan Kaufmann Publishers Inc.

Onishi, T., H. Wang, M. Bansal, K. Gimpel, and D. McAllester (2016). Who did what: A large-scale person-centered cloze dataset. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 2230–2235.

Owoputi, O., B. O'Connor, C. Dyer, K. Gimpel, N. Schneider, and N. A. Smith (2013). Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 380–390.

Packard, W., E. M. Bender, J. Read, S. Oepen, and R. Dridan (2014). Simple negation scope resolution through deep parsing: A semantic solution to a semantic problem. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 69–78.

Paice, C. D. (1990). Another stemmer. In *ACM SIGIR Forum*, Volume 24, pp. 56–61.

Pak, A. and P. Paroubek (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *LREC*, Volume 10, pp. 1320–1326.

Palmer, F. R. (2001). *Mood and modality*. Cambridge University Press.

Palmer, M., D. Gildea, and P. Kingsbury (2005). The proposition bank: An annotated corpus of semantic roles. *Computational linguistics 31*(1), 71–106.

Pan, S. J. and Q. Yang (2010). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering 22*(10), 1345–1359.

Pan, X., T. Cassidy, U. Hermjakob, H. Ji, and K. Knight (2015). Unsupervised entity linking with abstract meaning representation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 1130–1139.

Pang, B. and L. Lee (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 271–278.

Pang, B. and L. Lee (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 115–124.

Pang, B. and L. Lee (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval 2*(1-2), 1–135.

Pang, B., L. Lee, and S. Vaithyanathan (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 79–86.

Parsons, T. (1990). *Events in the Semantics of English*, Volume 5. MIT Press.

Pascanu, R., T. Mikolov, and Y. Bengio (2013). On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pp. 1310–1318.

Pedersen, T., S. Patwardhan, and J. Michelizzi (2004). Wordnet::similarity - measuring the relatedness of concepts. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 38–41.

Pei, W., T. Ge, and B. Chang (2015). An effective neural network model for graph-based dependency parsing. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 313–322.

Peng, F., F. Feng, and A. McCallum (2004). Chinese segmentation and new word detection using conditional random fields. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pp. 562.

Pennington, J., R. Socher, and C. Manning (2014). Glove: Global vectors for word representation. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 1532–1543.

Pereira, F. and Y. Schabes (1992). Inside-outside reestimation from partially bracketed corpora. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 128–135.

Pereira, F. C. N. and S. M. Shieber (2002). *Prolog and natural-language analysis*. Microtome Publishing.

Peterson, W. W., T. G. Birdsall, and W. C. Fox (1954). The theory of signal detectability. *Transactions of the IRE professional group on information theory 4*(4), 171–212.

Petrov, S., L. Barrett, R. Thibaux, and D. Klein (2006). Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Petrov, S., D. Das, and R. McDonald (2012, May). A universal part-of-speech tagset. In *Proceedings of LREC*.

Petrov, S. and R. McDonald (2012). Overview of the 2012 shared task on parsing the web. In *Notes of the First Workshop on Syntactic Analysis of Non-Canonical Language (SANCL)*, Volume 59.

Pinter, Y., R. Guthrie, and J. Eisenstein (2017). Mimicking word embeddings using subword RNNs. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.

Plank, B., A. Søgaard, and Y. Goldberg (2016). Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Polanyi, L. and A. Zaenen (2006). Contextual valence shifters. In *Computing attitude and affect in text: Theory and applications*. Springer.

Ponzetto, S. P. and M. Strube (2006). Exploiting semantic role labeling, wordnet and wikipedia for coreference resolution. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 192–199.

Ponzetto, S. P. and M. Strube (2007). Knowledge derived from wikipedia for computing semantic relatedness. *Journal of Artificial Intelligence Research 30*, 181–212.

Poon, H. and P. Domingos (2008). Joint unsupervised coreference resolution with markov logic. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 650–659.

Poon, H. and P. Domingos (2009). Unsupervised semantic parsing. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 1–10.

Popel, M., D. Marecek, J. Stepánek, D. Zeman, and Z. Zabokrtskỳ (2013). Coordination structures in dependency treebanks. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 517–527.

Popescu, A.-M., O. Etzioni, and H. Kautz (2003). Towards a theory of natural language interfaces to databases. In *Proceedings of Intelligent User Interfaces (IUI)*, pp. 149–157.

Poplack, S. (1980). Sometimes i'll start a sentence in spanish y termino en español: toward a typology of code-switching1. *Linguistics 18*(7-8), 581–618.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program 14*(3), 130–137.

Prabhakaran, V., O. Rambow, and M. Diab (2010). Automatic committed belief tagging. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pp. 1014–1022.

Pradhan, S., X. Luo, M. Recasens, E. Hovy, V. Ng, and M. Strube (2014). Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 30–35.

Pradhan, S., L. Ramshaw, M. Marcus, M. Palmer, R. Weischedel, and N. Xue (2011). CoNLL-2011 shared task: Modeling unrestricted coreference in OntoNotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pp. 1–27. Association for Computational Linguistics.

Pradhan, S., W. Ward, K. Hacioglu, J. H. Martin, and D. Jurafsky (2005). Semantic role labeling using different syntactic views. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 581–588.

Punyakanok, V., D. Roth, and W.-t. Yih (2008). The importance of syntactic parsing and inference in semantic role labeling. *Computational Linguistics 34*(2), 257–287.

Pustejovsky, J., P. Hanks, R. Sauri, A. See, R. Gaizauskas, A. Setzer, D. Radev, B. Sundheim, D. Day, L. Ferro, et al. (2003). The timebank corpus. In *Corpus linguistics*, Volume 2003, pp. 40. Lancaster, UK.

Pustejovsky, J., B. Ingria, R. Sauri, J. Castano, J. Littman, R. Gaizauskas, A. Setzer, G. Katz, and I. Mani (2005). The specification language timeml. In *The language of time: A reader*, pp. 545–557. Oxford University Press.

Qiu, G., B. Liu, J. Bu, and C. Chen (2011). Opinion word expansion and target extraction through double propagation. *Computational linguistics 37*(1), 9–27.

Quattoni, A., S. Wang, L.-P. Morency, M. Collins, and T. Darrell (2007). Hidden conditional random fields. *IEEE transactions on pattern analysis and machine intelligence 29*(10).

Rabiner, L. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE 77*(2), 257–286.

Rahman, A. and V. Ng (2011). Narrowing the modeling gap: a cluster-ranking approach to coreference resolution. *Journal of Artificial Intelligence Research 40*, 469–521.

Rajpurkar, P., J. Zhang, K. Lopyrev, and P. Liang (2016). Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 2383–2392.

Rao, D., D. Yarowsky, A. Shreevats, and M. Gupta (2010). Classifying latent user attributes in twitter. In *Proceedings of Workshop on Search and mining user-generated contents*.

Ratinov, L. and D. Roth (2009). Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pp. 147–155. Association for Computational Linguistics.

Ratinov, L., D. Roth, D. Downey, and M. Anderson (2011). Local and global algorithms for disambiguation to wikipedia. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 1375–1384.

Ratliff, N. D., J. A. Bagnell, and M. Zinkevich (2007). (approximate) subgradient methods for structured prediction. In *Proceedings of Artificial Intelligence and Statistics (AISTATS)*, pp. 380–387.

Ratnaparkhi, A. (1996). A maximum entropy model for part-of-speech tagging. In *emnlp*, pp. 133–142.

Ratnaparkhi, A., J. Reynar, and S. Roukos (1994). A maximum entropy model for prepositional phrase attachment. In *Proceedings of the workshop on Human Language Technology*, pp. 250–255. Association for Computational Linguistics.

Read, J. (2005). Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL student research workshop*, pp. 43–48. Association for Computational Linguistics.

Reisinger, D., R. Rudinger, F. Ferraro, C. Harman, K. Rawlins, and B. V. Durme (2015). Semantic proto-roles. *Transactions of the Association for Computational Linguistics 3*, 475–488.

Reisinger, J. and R. J. Mooney (2010). Multi-prototype vector-space models of word meaning. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 109–117.

Ribeiro, F. N., M. Araújo, P. Gonçalves, M. A. Gonçalves, and F. Benevenuto (2016). Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science 5*(1), 1–29.

Richardson, M., C. J. Burges, and E. Renshaw (2013). MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 193–203.

Riedel, S., L. Yao, and A. McCallum (2010). Modeling relations and their mentions without labeled text. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML)*, pp. 148–163.

Riloff, E. (1996). Automatically generating extraction patterns from untagged text. In *Proceedings of the national conference on artificial intelligence*, pp. 1044–1049.

Riloff, E. and J. Wiebe (2003). Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pp. 105–112. Association for Computational Linguistics.

Roark, B., M. Saraclar, and M. Collins (2007). Discriminative¡ i¿ n¡/i¿-gram language modeling. *Computer Speech & Language 21*(2), 373–392.

Robert, C. and G. Casella (2013). *Monte Carlo statistical methods*. Springer Science & Business Media.

Rosenfeld, R. (1996). A maximum entropy approach to adaptive statistical language modelling. *Computer Speech & Language 10*(3), 187–228.

Ross, S., G. Gordon, and D. Bagnell (2011). A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of Artificial Intelligence and Statistics (AISTATS)*, pp. 627–635.

Rush, A. M., D. Sontag, M. Collins, and T. Jaakkola (2010). On dual decomposition and linear programming relaxations for natural language processing. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 1–11.

Santos, C. D. and B. Zadrozny (2014). Learning character-level representations for part-of-speech tagging. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 1818–1826.

Sato, M.-A. and S. Ishii (2000). On-line em algorithm for the normalized gaussian network. *Neural computation 12*(2), 407–432.

Saurí, R. and J. Pustejovsky (2009). Factbank: a corpus annotated with event factuality. *Language resources and evaluation 43*(3), 227.

Saxe, A. M., J. L. McClelland, and S. Ganguli (2014). Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Schank, R. C. and R. Abelson (1977). *Scripts, goals, plans, and understanding*. Hillsdale, NJ: Erlbaum.

Schapire, R. E. and Y. Singer (2000). Boostexter: A boosting-based system for text categorization. *Machine learning 39*(2-3), 135–168.

Schnabel, T., I. Labutov, D. Mimno, and T. Joachims (2015). Evaluation methods for unsupervised word embeddings. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 298–307.

Schneider, N., J. Flanigan, and T. O'Gorman (2015). The logic of amr: Practical, unified, graph-based sentence semantics for nlp. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 4–5.

Schütze, H. (1998). Automatic word sense discrimination. *Computational linguistics 24*(1), 97–123.

Sennrich, R., B. Haddow, and A. Birch (2016). Neural machine translation of rare words with subword units. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 1715–1725.

Settles, B. (2012). Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning 6*(1), 1–114.

Shen, D. and M. Lapata (2007). Using semantic roles to improve question answering. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 12–21.

Shen, W., J. Wang, and J. Han (2015). Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering 27*(2), 443–460.

Shieber, S. M. (1985). Evidence against the context-freeness of natural language. *Linguistics and Philosophy 8*(3), 333–343.

Singh, S., A. Subramanya, F. Pereira, and A. McCallum (2011). Large-scale cross-document coreference using distributed inference and hierarchical models. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 793–803.

Sipser, M. (2012). *Introduction to the Theory of Computation*. Cengage Learning.

Smith, D. A. and J. Eisner (2008). Dependency parsing by belief propagation. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 145–156.

Smith, D. A. and N. A. Smith (2007). Probabilistic models of nonprojective dependency trees. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 132–140.

Smith, N. A. (2011). Linguistic structure prediction. *Synthesis Lectures on Human Language Technologies 4*(2), 1–274.

Snow, R., B. O'Connor, D. Jurafsky, and A. Y. Ng (2008). Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 254–263.

Socher, R., J. Bauer, C. D. Manning, and A. Y. Ng (2013). Parsing with compositional vector grammars. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Socher, R., B. Huval, C. D. Manning, and A. Y. Ng (2012). Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 1201–1211. Association for Computational Linguistics.

Socher, R., A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.

Søgaard, A. (2013). Semi-supervised learning and domain adaptation in natural language processing. *Synthesis Lectures on Human Language Technologies 6*(2), 1–103.

Solorio, T. and Y. Liu (2008). Learning to predict code-switching points. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 973–981. Association for Computational Linguistics.

Somasundaran, S. and J. Wiebe (2009). Recognizing stances in online debates. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 226–234.

Song, L., B. Boots, S. M. Siddiqi, G. J. Gordon, and A. J. Smola (2010). Hilbert space embeddings of hidden markov models. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 991–998.

Soon, W. M., H. T. Ng, and D. C. Y. Lim (2001). A machine learning approach to coreference resolution of noun phrases. *Computational linguistics 27*(4), 521–544.

Sowa, J. F. (2000). *Knowledge representation: logical, philosophical, and computational foundations*. Pacific Grove, CA: Brooks/Cole.

Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation 28*(1), 11–21.

Spitkovsky, V. I., H. Alshawi, D. Jurafsky, and C. D. Manning (2010). Viterbi training improves unsupervised dependency parsing. In *CONLL*, pp. 9–17.

Sproat, R., A. Black, S. Chen, S. Kumar, M. Ostendorf, and C. Richards (2001). Normalization of non-standard words. *Computer Speech & Language 15*(3), 287–333.

Sproat, R., W. Gale, C. Shih, and N. Chang (1996). A stochastic finite-state word-segmentation algorithm for chinese. *Computational linguistics 22*(3), 377–404.

Sra, S., S. Nowozin, and S. J. Wright (2012). *Optimization for machine learning*. MIT Press.

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research 15*(1), 1929–1958.

Srivastava, R. K., K. Greff, and J. Schmidhuber (2015). Training very deep networks. In *Neural Information Processing Systems (NIPS)*, pp. 2377–2385.

Steedman, M. and J. Baldridge (2011). Combinatory categorial grammar. In *Non-Transformational Syntax: Formal and Explicit Models of Grammar*. Wiley-Blackwell.

Stenetorp, P., S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii (2012). Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 102–107.

Stern, M., J. Andreas, and D. Klein (2017). A minimal span-based neural constituency parser. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Stolcke, A., K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics 26*(3), 339–373.

Stone, P. J. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. The MIT Press.

Stoyanov, V., N. Gilbert, C. Cardie, and E. Riloff (2009). Conundrums in noun phrase coreference resolution: Making sense of the state-of-the-art. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 656–664.

Strubell, E., P. Verga, D. Belanger, and A. McCallum (2017). Fast and accurate entity recognition with iterated dilated convolutions. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.

Suchanek, F. M., G. Kasneci, and G. Weikum (2007). Yago: a core of semantic knowledge. In *Proceedings of the Conference on World-Wide Web (WWW)*, pp. 697–706.

Sun, X., T. Matsuzaki, D. Okanohara, and J. Tsujii (2009). Latent variable perceptron algorithm for structured classification. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, Volume 9, pp. 1236–1242.

Sun, Y., L. Lin, D. Tang, N. Yang, Z. Ji, and X. Wang (2015). Modeling mention, context and entity with neural networks for entity disambiguation. In *IJCAI*, pp. 1333–1339.

Sundermeyer, M., R. Schlüter, and H. Ney (2012). Lstm neural networks for language modeling. In *INTERSPEECH*.

Surdeanu, M., J. Tibshirani, R. Nallapati, and C. D. Manning (2012). Multi-instance multi-label learning for relation extraction. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 455–465.

Sutton, R. S. and A. G. Barto (1998). *Reinforcement learning: An introduction*, Volume 1. MIT press Cambridge.

Sutton, R. S., D. A. McAllester, S. P. Singh, and Y. Mansour (2000). Policy gradient methods for reinforcement learning with function approximation. In *Neural Information Processing Systems (NIPS)*, pp. 1057–1063.

Taboada, M., J. Brooke, M. Tofiloski, K. Voll, and M. Stede (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics 37*(2), 267–307.

Täckström, O., K. Ganchev, and D. Das (2015). Efficient inference and structured learning for semantic role labeling. *Transactions of the Association for Computational Linguistics 3*, 29–41.

Täckström, O., R. McDonald, and J. Uszkoreit (2012). Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 477–487.

Tang, D., B. Qin, and T. Liu (2015). Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 1422–1432.

Tarjan, R. E. (1977). Finding optimum branchings. *Networks 7*(1), 25–35.

Taskar, B., C. Guestrin, and D. Koller (2003). Max-margin markov networks. In *Neural Information Processing Systems (NIPS)*.

Tausczik, Y. R. and J. W. Pennebaker (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology 29*(1), 24–54.

Teh, Y. W. (2006). A hierarchical bayesian language model based on pitman-yor processes. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 985–992.

Tesnière, L. (1966). *Éléments de syntaxe structurale* (second ed.). Paris: Klincksieck.

Thomas, M., B. Pang, and L. Lee (2006). Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 327–335.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Toutanova, K., D. Klein, C. D. Manning, and Y. Singer (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Tromble, R. W. and J. Eisner (2006). A fast finite-state relaxation method for enforcing global constraints on sequence decoding. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 423.

Tsochantaridis, I., T. Hofmann, T. Joachims, and Y. Altun (2004). Support vector machine learning for interdependent and structured output spaces. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 104. ACM.

Tsvetkov, Y., M. Faruqui, W. Ling, G. Lample, and C. Dyer (2015). Evaluation of word vector representations by subspace alignment. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 2049–2054.

Turian, J., L. Ratinov, and Y. Bengio (2010). Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 384–394.

Turney, P. D. and P. Pantel (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research 37*, 141–188.

Twain, M. (1997). *A Tramp Abroad*. New York: Penguin.

Tzeng, E., J. Hoffman, T. Darrell, and K. Saenko (2015). Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4068–4076.

Usunier, N., D. Buffoni, and P. Gallinari (2009). Ranking with ordered weighted pairwise classification. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 1057–1064.

Uzuner, Ö., X. Zhang, and T. Sibanda (2009). Machine learning and rule-based approaches to assertion classification. *Journal of the American Medical Informatics Association 16*(1), 109–115.

Vadas, D. and J. R. Curran (2011). Parsing noun phrases in the penn treebank. *Computational Linguistics 37*(4), 753–809.

Van Eynde, F. (2006). NP-internal agreement and the structure of the noun phrase. *Journal of Linguistics 42*(1), 139–186.

Van Gael, J., A. Vlachos, and Z. Ghahramani (2009). The infinite hmm for unsupervised pos tagging. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 678–687.

Velldal, E., L. Øvrelid, J. Read, and S. Oepen (2012). Speculation and negation: Rules, rankers, and the role of syntax. *Computational linguistics 38*(2), 369–410.

Vilain, M., J. Burger, J. Aberdeen, D. Connolly, and L. Hirschman (1995). A model-theoretic coreference scoring scheme. In *Proceedings of the 6th conference on Message understanding*, pp. 45–52. Association for Computational Linguistics.

Vincent, P., H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research 11*(Dec), 3371–3408.

Vincze, V., G. Szarvas, R. Farkas, G. Móra, and J. Csirik (2008). The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC bioinformatics 9*(11), S9.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory 13*(2), 260–269.

Wager, S., S. Wang, and P. S. Liang (2013). Dropout training as adaptive regularization. In *Neural Information Processing Systems (NIPS)*, pp. 351–359.

Wainwright, M. J. and M. I. Jordan (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning 1*(1-2), 1–305.

Walker, M. A. (2000). An application of reinforcement learning to dialogue strategy selection in a spoken dialogue system for email. *Journal of Artificial Intelligence Research 12*, 387–416.

Wang, C., N. Xue, and S. Pradhan (2015). A Transition-based Algorithm for AMR Parsing. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 366–375.

Wang, H., T. Onishi, K. Gimpel, and D. McAllester (2017). Emergent predication structure in hidden state vectors of neural readers. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pp. 26–36.

Weaver, W. (1955). Translation. *Machine translation of languages 14*, 15–23.

Wei, G. C. and M. A. Tanner (1990). A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association 85*(411), 699–704.

Weinberger, K., A. Dasgupta, J. Langford, A. Smola, and J. Attenberg (2009). Feature hashing for large scale multitask learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 1113–1120.

Weston, J., S. Bengio, and N. Usunier (2011). Wsabie: Scaling up to large vocabulary image annotation. In *IJCAI*, Volume 11, pp. 2764–2770.

Wiebe, J., T. Wilson, and C. Cardie (2005). Annotating expressions of opinions and emotions in language. *Language resources and evaluation 39*(2), 165–210.

Wieting, J., M. Bansal, K. Gimpel, and K. Livescu (2015). Towards universal paraphrastic sentence embeddings. *arXiv preprint arXiv:1511.08198*.

Wieting, J., M. Bansal, K. Gimpel, and K. Livescu (2016). CHARAGRAM: Embedding words and sentences via character n-grams. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 1504–1515.

Wilson, T., J. Wiebe, and P. Hoffmann (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 347–354.

Winograd, T. (1972). Understanding natural language. *Cognitive psychology 3*(1), 1–191.

Wiseman, S., A. M. Rush, and S. M. Shieber (2016). Learning global features for coreference resolution. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 994–1004.

Wiseman, S. J., A. M. Rush, S. M. Shieber, and J. Weston (2015). Learning anaphoricity and antecedent ranking features for coreference resolution. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Wolfe, T., M. Dredze, and B. Van Durme (2017). Pocket knowledge base population. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 305–310.

Wong, Y. W. and R. J. Mooney (2006). Learning for semantic parsing with statistical machine translation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 439–446.

Wu, B. Y. and K.-M. Chao (2004). *Spanning trees and optimization problems*. CRC Press.

Wu, F. and D. S. Weld (2010). Open information extraction using wikipedia. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 118–127.

Xia, F. (2000). The part-of-speech tagging guidelines for the penn chinese treebank (3.0). Technical report, University of Pennsylvania Institute for Research in Cognitive Science.

Xu, W., X. Liu, and Y. Gong (2003). Document clustering based on non-negative matrix factorization. In *SIGIR*, pp. 267–273. ACM.

Xu, Y., L. Mou, G. Li, Y. Chen, H. Peng, and Z. Jin (2015). Classifying relations via long short term memory networks along shortest dependency paths. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 1785–1794.

Xue, N. et al. (2003). Chinese word segmentation as character tagging. *Computational Linguistics and Chinese Language Processing 8*(1), 29–48.

Yamada, H. and Y. Matsumoto (2003). Statistical dependency analysis with support vector machines. In *Proceedings of IWPT*, Volume 3, pp. 195–206.

Yang, Y., M.-W. Chang, and J. Eisenstein (2016). Toward socially-infused information extraction: Embedding authors, mentions, and entities. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.

Yang, Y. and J. Eisenstein (2013). A log-linear model for unsupervised text normalization. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.

Yang, Y. and J. Eisenstein (2015). Unsupervised multi-domain adaptation with feature embeddings. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Yang, Y., W.-t. Yih, and C. Meek (2015). WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 2013–2018.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 189–196. Association for Computational Linguistics.

Yee, L. C. and T. Y. Jones (2012, March). Apple ceo in china mission to clear up problems. *Reuters*. retrieved on March 26, 2017.

Yu, C.-N. J. and T. Joachims (2009). Learning structural svms with latent variables. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 1169–1176.

Yu, F. and V. Koltun (2016). Multi-scale context aggregation by dilated convolutions. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Zeiler, M. D. (2012). Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

Zelenko, D., C. Aone, and A. Richardella (2003). Kernel methods for relation extraction. *The Journal of Machine Learning Research 3*, 1083–1106.

Zelle, J. M. and R. J. Mooney (1996). Learning to parse database queries using inductive logic programming. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, pp. 1050–1055.

Zeng, D., K. Liu, S. Lai, G. Zhou, and J. Zhao (2014). Relation classification via convolutional deep neural network. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pp. 2335–2344.

Zettlemoyer, L. S. and M. Collins (2005). Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In *Proceedings of UAI*.

Zhang, T. (2004). Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 116. ACM.

Zhang, X., J. Zhao, and Y. LeCun (2015). Character-level convolutional networks for text classification. In *Neural Information Processing Systems (NIPS)*, pp. 649–657.

Zhang, Y. and S. Clark (2008). A tale of two parsers: investigating and combining graph-based and transition-based dependency parsing using beam-search. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 562–571.

Zhang, Y., T. Lei, R. Barzilay, T. Jaakkola, and A. Globerson (2014). Steps to excellence: Simple inference with refined scoring of dependency trees. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 197–207.

Zhang, Y. and J. Nivre (2011). Transition-based dependency parsing with rich non-local features. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 188–193.

Zhou, J. and W. Xu (2015). End-to-end learning of semantic role labeling using recurrent neural networks. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 1127–1137.

Zhu, J., Z. Nie, X. Liu, B. Zhang, and J.-R. Wen (2009). Statsnowball: a statistical approach to extracting entity relationships. In *Proceedings of the Conference on World-Wide Web (WWW)*, pp. 101–110.

Zhu, X., Z. Ghahramani, and J. D. Lafferty (2003). Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 912–919.

Zhu, X. and A. B. Goldberg (2009). Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning 3*(1), 1–130.

Zou, W. Y., R. Socher, D. Cer, and C. D. Manning (2013). Bilingual word embeddings for phrase-based machine translation. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 1393–1398.

# Index