

Metadata of the book that will be visualized in SpringerLink

Publisher Name	Springer Singapore	
Publisher Location	Singapore	
Series ID	7092	
SeriesTitle	Studies in Computational Intelligence	
Book ID	454313_1_En	
Book Title	Neural Representations of Natural Language	
Book DOI	10.1007/978-981-13-0062-2	
Copyright Holder Name	Springer Nature Singapore Pte Ltd.	
Copyright Year	2019	
Author	Family Name	White
	Particle	
	Given Name	Lyndon
	Suffix	
	Division	School of Electrical Electronics and Computer Engineering
	Organization	The University of Western Australia
	Address	Perth, Western Autralia, Australia
	Email	LYNDON.WHITE@RESEARCH.UWA.EDU.AU
Author	Family Name	Togneri
	Particle	
	Given Name	Roberto
	Suffix	
	Division	School of Electrical Electronics and Computer Engineering
	Organization	The University of Western Australia
	Address	Crawley, Western Autralia, Australia
	Email	Roberto.Togneri@uwa.edu.au
Author	Family Name	Liu
	Particle	
	Given Name	Wei
	Suffix	
	Division	School of Computer Science and Software Engineering
	Organization	The University of Western Australia
	Address	Perth, Western Autralia, Australia
	Email	wei.liu@uwa.edu.au
Author	Family Name	Bennamoun
	Particle	
	Given Name	Mohammed
	Suffix	
	Division	School of Computer Science and Software Engineering
	Organization	The University of Western Australia
	Address	Perth, Western Autralia, Australia

Email

MOHAMMED.BENNAMOUN@UWA.EDU.AU



1
2

Studies in Computational Intelligence

3
4

Volume 783

5

Series editor

6
7
8

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland
e-mail: kacprzyk@ibspan.waw.pl



9 The series “Studies in Computational Intelligence” (SCI) publishes new developments and advances in the various areas of computational intelligence—quickly and
10 with a high quality. The intent is to cover the theory, applications, and design methods of computational intelligence, as embedded in the fields of engineering,
11 computer science, physics and life sciences, as well as the methodologies behind them. The series contains monographs, lecture notes and edited volumes in
12 computational intelligence spanning the areas of neural networks, connectionist systems,
13 genetic algorithms, evolutionary computation, artificial intelligence, cellular automata, self-organizing systems, soft computing, fuzzy systems, and hybrid intelligent systems. Of particular value to both the contributors and the
14 readership are the short publication timeframe and the world-wide distribution,
15 which enable both wide and rapid dissemination of research output.
16

21 More information about this series at <http://www.springer.com/series/7092>

22



23

24
25

Lyndon White • Roberto Togneri
Wei Liu • Mohammed Bennamoun

26
27

Neural Representations of Natural Language

28





50 Lyndon White
51 School of Electrical, Electronic
52 and Computer Engineering
53 The University of Western Australia
54 Perth, WA, Australia

55 Roberto Tognier
56 School of Electrical, Electronic
57 and Computer Engineering
58 The University of Western Australia
59 Perth, WA, Australia

Wei Liu
School of Computer Science
and Software Engineering
The University of Western Australia
Perth, WA, Australia

Mohammed Bennamoun
School of Computer Science
and Software Engineering
The University of Western Australia
Perth, WA, Australia

50
51 LW: University has restructured recently. Ideally would use new affiliations
52

Dept. of Computer Science and Software Engineering
School of Physics, Mathematics and Computing
Faculty of Engineering and Mathematical Sciences
The University of Western Australia

Dept. of Electrical, Electronic and Computer Engineering
School of Engineering
Faculty of Engineering and Mathematical Sciences
The University of Western Australia

53

54 ISSN 1860-949X ISSN 1860-9503 (electronic)
55 Studies in Computational Intelligence
56 ISBN 978-981-13-0061-5 ISBN 978-981-13-0062-2 (eBook)
57 https://doi.org/10.1007/978-981-13-0062-2

60

61 Library of Congress Control Number: 2018938363
62

63 © Springer Nature Singapore Pte Ltd. 2019

64 This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part
65 of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations,
66 recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission
67 or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar
68 methodology now known or hereafter developed.

69 The use of general descriptive names, registered names, trademarks, service marks, etc. in this
70 publication does not imply, even in the absence of a specific statement, that such names are exempt from
71 the relevant protective laws and regulations and therefore free for general use.

72 The publisher, the authors and the editors are safe to assume that the advice and information in this
73 book are believed to be true and accurate at the date of publication. Neither the publisher nor the
74 authors or the editors give a warranty, express or implied, with respect to the material contained herein or
75 for any errors or omissions that may have been made. The publisher remains neutral with regard to
76 jurisdictional claims in published maps and institutional affiliations.

77 This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
78 The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721,
79 Singapore



80

Preface

81

82

83

84 *We do, therefore, offer you our practical judgements wherever we can. As you gain*
85 *experience, you will form your own opinion of how reliable our advice is. Be assured that it*
86 *is not perfect!*

87 —Press, Teukolsky, Vettering and Flannery, *Numerical Recipes*, 3rd ed., 2007

88 Some might wonder why one would be concerned with finding good representations
89 for natural language. To answer that simply, all problem-solving is significantly simplified by finding a good representation, whether that is reducing a
90 real-world problem into a system of inequalities to be solved by a constrained
91 optimizer or using domain-specific jargon to communicate with experts in the area.
92

A wealth of other materials

Wikipedia articles on ML and NLP tend to be of reasonable quality, Coursera offers several courses on the topics, Cross Validated Stack Exchange has thousands of QAs, and the documentation for many machine learning libraries often contains high-quality tutorials. Finally, preprints of the majority of the papers in the fields are available on arXiv.

93 There is a truly immense amount of existing information, designed for consumption
94 by humans. Much of it is in text form. People prefer to create and consume
95 information in natural language (such as prose) format, rather than in some more
96 computationally accessible format (such as a directed graph). For example, doctors
97 prefer to dictate clinical notes over filling in forms for a database.

98 One of the core attractions of a deep learning system is that it functions by
99 learning increasingly abstract representations of inputs. These abstract representations
100 include useful features for the task. Importantly, they also include features
101 applicable even for more distantly related tasks.

102 It is the general assumption of this book that it is not being read in isolation. That
103 the reader is not bereft of all other sources of knowledge. We assume not only can
104 the reader access the primary sources for the papers we cite, but also that they are
105 able to discover and access other suitable materials, in order to go deeper on related



areas. There exist a wealth of blog posts, video tutorials, encyclopaedic article, etc., on machine learning and on the mathematics involved.

In general, we do assume the reader is familiar with matrix multiplication. In general, we will define networks in terms of matrices (rather than sums), as this is more concise, and better reflects real-world applications, in terms of code that a programmer would write. We also assume a basic knowledge of probability, and an even more basic knowledge of English linguistics. Though they are not the intended audience, very little of the book should be beyond someone with a high-school education.

The core focus of this book is to summarize the techniques that have been developed over the past 1–2 decades. In doing so, we detail the works of several dozen papers (and cite well over a hundred). Significant effort has gone into describing these works clearly and consistently. In doing so, the notation used does not normally line up with the original notations used in those papers; however, we ensure it is always mathematically equivalent. For some techniques, the explanation is short and simple. For other more challenging ideas, our explanation may be much longer than the original very concise formulation that is common in some papers.

For brevity, we have had to limit discussion of some aspect of natural language. In particular, we have neglected all discussion on the notion that a word may be made up of multiple tokens, for example, made up of. Phrases do receive some discussion in Chap. 6. Works such as Yin and Schütze (2014) exploration deserve more attention than we have space to give them.

Similarly, we do not have space to cover character-based models, mainly character RNNs, and other deep character models such as Zhang and LeCun (2015). These models have relevance both as sources from which word representations can be derived (Bojanowski et al. 2017), but more generally can be used for end-to-end systems. Using a purely character-based approach forces the model to learn tokenizing, parsing, and any other feature extraction internally. We focus instead on models that work within larger systems that accomplish such preprocessing.

Finally, we omit all discussion of attention models in recurrent neural networks (Bahdanau et al. 2014). The principles to the application remain similar to using normal recurrent neural networks in the applications discussed in this book. The additional attention features allow several improvements, making it the state of the art for many sequential tasks. We trust the reader with the experience of such models and will see how they may be applied to the uses of RNNs discussed here.

This book makes extensive use of text in grey boxes. This is used to provide reminders of definitions and comments on potentially unfamiliar notations. As well as to highlight non-technical aspects (such as societal concerns), and conversely overly technical aspects (such as implementation details) of the works discussed. They are also used to give the titles to citations (to save flicking to the huge list of references at the back of the book) and for the captions to the figures and more generally to provide non-essential but potentially helpful information without interrupting the flow of the main text. One could read the whole book without ever

AQ2

AQ3

LW: In the other sections of the book the citations (underlined) are hyperlinks to the bibliography.

The font in "made up of". It should be typewriter text (matching other natural language elements), not sans.



149 reading any of the text in grey boxes, as they are not required to understand the
150 main text. However, one would be missing out on a portion of, some very inter-
151 esting, content.

152 Perth, Australia

153

154

155

156

Lyndon White
Roberto Tognieri
Wei Liu
Mohammed Bennamoun

157

References

- 159 Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by
160 jointly learning to align and translate. In *CoRR* abs/1409.0473. [arXiv:1409.0473](https://arxiv.org/abs/1409.0473).
- 161 Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word
162 vectors with subword information. *Transactions of the Association for Computational
163 Linguistics*, vol. 5, 135–146.
- 164 Yin, Wenpeng and Hinrich Schütze. 2014. An exploration of embeddings for generalized phrases.
165 In *ACL* 2014, p. 41.
- 166 Zhang, Xiang and Yann Le Cun. 2015. Text understanding from scratch. In *CoRR proceedings
167 of the 52nd annual meeting of the association for computational linguistics*, vol. 11.



Contents

172	1 Introduction to Neural Networks for Machine Learning	1
173	1.1 Neural Networks	2
175	1.2 Function Approximation	4
178	1.3 Network Hyper-Parameters	4
180	1.3.1 Width and Depth	5
182	1.3.2 Activation Functions	5
184	1.4 Training	10
186	1.4.1 Gradient Descent and Back-Propagation	12
188	1.5 Some Examples of Common Neural Network Architectures	14
190	1.5.1 Classifier	14
192	1.5.2 Bottlenecking Autoencoder	17
194	References	21
196	2 Recurrent Neural Networks for Sequential Processing	23
198	2.1 Recurrent Neural Networks	24
200	2.2 General RNN Structures	24
202	2.2.1 Matched-Sequence	26
204	2.2.2 Encoder	26
206	2.2.3 Decoder	27
208	2.2.4 Encoder-Decoder	28
210	2.3 Inside the Recurrent Unit	29
212	2.3.1 Basic Recurrent Unit	29
214	2.3.2 Gated Recurrent Unit	30
216	2.3.3 LSTM Recurrent Unit	32
218	2.4 Further Variants	34
220	2.4.1 Deep Variants	34
222	2.4.2 Bidirectional RNNs	34
224	2.4.3 Other RNNs	34
226	References	35



228	3 Word Representations	37
230	3.1 Representations for Language Modeling	39
232	3.1.1 The Neural Probabilistic Language Model	41
234	3.1.2 RNN Language Models	47
236	3.2 Acausal Language Modeling	49
238	3.2.1 Continuous Bag of Words	50
240	3.2.2 Skip-Gram	51
242	3.2.3 Analogy Tasks	53
244	3.3 Co-location Factorisation	55
246	3.3.1 GloVe	55
247	3.3.2 Further Equivalence of Co-location Prediction to Factorisation	57
249	3.3.3 Conclusion	57
250	3.4 Hierarchical Softmax and Negative Sampling	57
253	3.4.1 Hierarchical Softmax	57
255	3.4.2 Negative Sampling	64
258	3.5 Natural Language Applications – Beyond Language Modeling	68
260	3.5.1 Using Word Embeddings as Features	68
262	3.6 Aligning Vector Spaces Across Languages	69
264	References	70
268	4 Word Sense Representations	73
269	4.1 Word Senses	74
272	4.1.1 Word Sense Disambiguation	75
274	4.2 Word Sense Representation	79
276	4.2.1 Directly Supervised Method	79
278	4.2.2 Word Embedding-Based Disambiguation Method	79
280	4.3 Word Sense Induction (WSI)	82
282	4.3.1 Context Clustering-Based Approaches	83
284	4.3.2 Co-location Prediction-Based Approaches	86
286	4.4 Conclusion	90
288	References	90
290	5 Sentence Representations and Beyond	93
292	5.1 Unordered and Weakly Ordered Representations	94
294	5.1.1 Sums of Word Embeddings	94
296	5.1.2 Paragraph Vector Models (Defunct)	96
298	5.2 Sequential Models	97
300	5.2.1 VAE and Encoder-Decoder	98
302	5.2.2 Skip-Thought	100



Contents

xi

304	5.3 Structured Models	100
305	5.3.1 Constituency Parse Tree (Binary)	104
308	5.3.2 Dependency Tree	105
309	5.3.3 Parsing	107
312	5.3.4 Recursive Autoencoders	108
314	5.4 Matrix Vector Models	110
316	5.4.1 Structured Matrix Vector Model	110
318	5.4.2 Sequential Matrix Vector Model	111
320	5.5 Conclusion, on Compositionality	113
322	References	114
324	6 Conclusion	117
326	References	121
328	Index	123
329		



331

332

Notations

333 The following notation is used throughout this work.

334

335	a	A scalar (real, integer, or word/token)
336	\tilde{a}	A vector, nominally a column vector
337	A	A matrix
338	\mathcal{A}	A sequence, including a data set or a sequence of words
339	\mathbb{V}	A set, e.g. the vocabulary
340	\tilde{x}_i	The i th element of the vector \tilde{x}
341	$X_{i,j}$	The row i and column j element X
342	X_i	The i th column <i>vector</i> of the matrix X
343	$X_{:,i}$	The i th row <i>vector</i> of the matrix X
344	w^t	A scalar t th element of some sequence
345	W^f	A matrix disambiguated by the name f
346	$[A \ B]$	The horizontal concatenation of A and B
347	$[A; B]$	The vertical concatenation of A and B
348	$P(\dots)$	A probability (estimated or ground truth)
349	A	A random variable (when not a matrix)
350	\hat{y}	A network output value, corresponding to target value y or \tilde{y} a vector or scalar quantity as appropriate

Words are treated as integers

We consistently notate words, as if they were scalar integer values. writing, for example, w^1 as to be the first word in a sequence, which is then used as an index: $C_{:,w^1}$ is its corresponding word vector, from the embedding matrix C .

351

352

353

354

355



Superscripts and Subscripts

Readers may wonder why we are using x_i , and x^i . Would not x_i suffice? Why differentiate between elements of a sequence and elements of a vector?

The particularly problematic case is that we often want to represent taking the i th element of a vector that is the t th element of a sequence of vectors. The vector, we would call \tilde{x}^t , its i element is \tilde{x}_i^t .

This is also not ambiguous with the matrix indexing notation $X_{i,t}$.

Rarely, a superscript will be actually an exponent, e.g. $x^{\frac{2}{3}}$. This should be apparent when in this case. Far more common is the natural exponent which we write $\exp(x)$.

356

357

LW: Is it possible to fit this box onto the previous page?

Author Query Form

359

361

362

Book ID : **454313_1_En**Chapter No : **7092**

the language of science

363

Please ensure you fill out your response to the queries raised below and return this form along with your corrections.

364

365

366

367

368

369

Dear Author,

During the process of typesetting your chapter, the following queries have arisen. Please check your typeset proof carefully against the queries listed below and mark the necessary changes either directly on the proof/online grid or in the ‘Author’s response’ area provided below

374

375

376

377

378

379

380

381

382

383

384

385

386

387

Query Refs.	Details Required	Author's Response
AQ1	Please check and confirm if the authors affiliations have been correctly identified. Amend if necessary.	
AQ2	‘The marginal text styling’ has been changed into ‘gray box styling’, according to which the paragraph ‘This book makes extensive ...’ has been modified. Please check and confirm.	
AQ3	Please check the clarity of the sentence ‘As well as to highlight non-technical aspects....’.	

388

389

LW: AQ1: It is "School of Electrical, Electronic and Computer Engineering", not "Electronics"

AQ1, AQ2 seems fine to me.

Other issues marked inline on pages

Metadata of the chapter that will be visualized in SpringerLink

Book Title	Neural Representations of Natural Language	
Series Title		
Chapter Title	Introduction to Neural Networks for Machine Learning	
Copyright Year	2019	
Copyright HolderName	Springer Nature Singapore Pte Ltd.	
Corresponding Author	Family Name Particle Given Name Prefix Suffix Role Division Organization Address Email	White Lyndon Togneri Roberto Liu Wei Bennamoun Mohammed
Author	Division Organization Address Email	School of Electrical, Electronic and Computer Engineering The University of Western Australia Perth, WA, Australia LYNDON.WHITE@RESEARCH.UWA.EDU.AU
Author	Division Organization Address Email	School of Electrical, Electronic and Computer Engineering The University of Western Australia Perth, WA, Australia
Author	Division Organization Address Email	School of Computer Science and Software Engineering The University of Western Australia Perth, WA, Australia
Author	Division Organization Address Email	School of Computer Science and Software Engineering The University of Western Australia Perth, WA, Australia

Role
Division School of Computer Science and Software Engineering
Organization The University of Western Australia
Address Perth, WA, Australia
Email

Abstract	This chapter covers the crucial machine learning techniques required to understand the remained of the book: namely neural networks. Readers already familiar with neural networks can freely skip this chapter. Readers interested in a more comprehensive coverage of all aspects of machine learning are referred to the many textbooks on this subject matter.
----------	--



Introduction to Neural Networks for Machine Learning

LW: Where is the chapter abstract gone?

LW: Previous epigraph was set as full-width. This one is half width.
They should be consistant.

Whether one sees the artificial neural network technique described below as learning or as optimization depends largely on one's background and one's theoretical likes and dislikes. I will freely use "learning" in the remainder of this section because that term is traditionally used in the neural networks literature, but the reader should feel free to substitute "optimization" if (s)he finds the other term offensive. Please contact the author for an Emacs lisp function to enforce the usage of choice

J. M. G. Lammens, PhD Dissertation, A Computational Model of Color Perception and Color Naming, State University of New York, 1994

Suggested Reading

As this is not a comprehensive introduction we recommend the reader to look elsewhere for additional reading. In particular we recommend the free online web-book: Nielsen and Michael (2015), and the comprehensive: Goodfellow et al. (2016).

This chapter covers the crucial machine learning techniques required to understand the remained of the book: namely neural networks. Readers already familiar with neural networks can freely skip this chapter. Readers interested in a more comprehensive coverage of all aspects of machine learning are referred to the many textbooks on this subject matter.

Machine learning, particularly neural networks, has become a hot topic in recent years. The core notion of machine learning is to learn to perform a function based on examples. This is in-contrast to "regular programming" where code is written to accomplish the function based on the programmer's analysis of the task. Machine



learning is at its most useful when it is difficult to articulate explicit rules for finding an output from a given input; but for which there are many examples of such. For example, the rule of English spelling “I before E, except after C”, is known to be often incorrect. It is hard to write a rule about letter order. However, a dictionary (or any other corpus) will have thousands of words showing the correct order. By applying suitable machine learning techniques, a system could learn to determine the correct spelling of words containing *i e* and *e i*, and inform the users as to if an input word is correct, even if the word never occurred in the training dictionary. This is because the machine learning algorithm (ideally) discovers generalisable rules, from the training examples. For most machine learning methods, including neural networks, these rules are not in a readily interpretable form, but are stored as numerical parameters of the model. It is for this reason they are often called black-box models.

21 1.1 Neural Networks

Multilayer perceptron or Artificial Neural Network?

An artificial neural network is sometimes called a multilayer perceptron. This is in reference to the (distantly) related perceptron learning algorithm. This term has generally fallen out of favour in newer works, with Geoffrey Hinton, who originally coined the term, expressing his own regret at the naming. Sometimes the term multilayer perceptron is used to distinguish feed-forward neural networks, from other neural network related technologies, such as restricted Boltzmann machines, Hopfield networks, self-organizing maps etc. Unless otherwise specified, we use the term neural network to refer to a feed-forward network as discussed in this chapter.

22 A neural network is one particular family of machine learning algorithms. It is
23 important to understand that neural networks are not emulations of the brain, they
24 are algorithms that were inspired by the ideas about how the brain worked (Hebb
25 1949). Some advancements have also been inspired by the functioning of the brain,
26 but many others have come from statistical methods or elsewhere. A neural network
27 is no more similar to a brain, than it is to a Fourier transform.

28 The core idea behind a neural network is to represent the transformation of the
29 input to the output as links between neurons in a sequence of layers. This is shown
30 in Fig. 1.1. Each neuron has a weighted connection to the neurons in the layer below,
31 and it applies a nonlinear function to this weighted sum, to determine its own output.
32 The output is connected to the inputs in the layer above, or is the final output.

Don't implement neurons

Typical object orientated programming teaches one to look for objects. A
obvious object in a neural network would be a neuron. Another object could

LW: Avoid breaking the grey box across 2 pages with image between

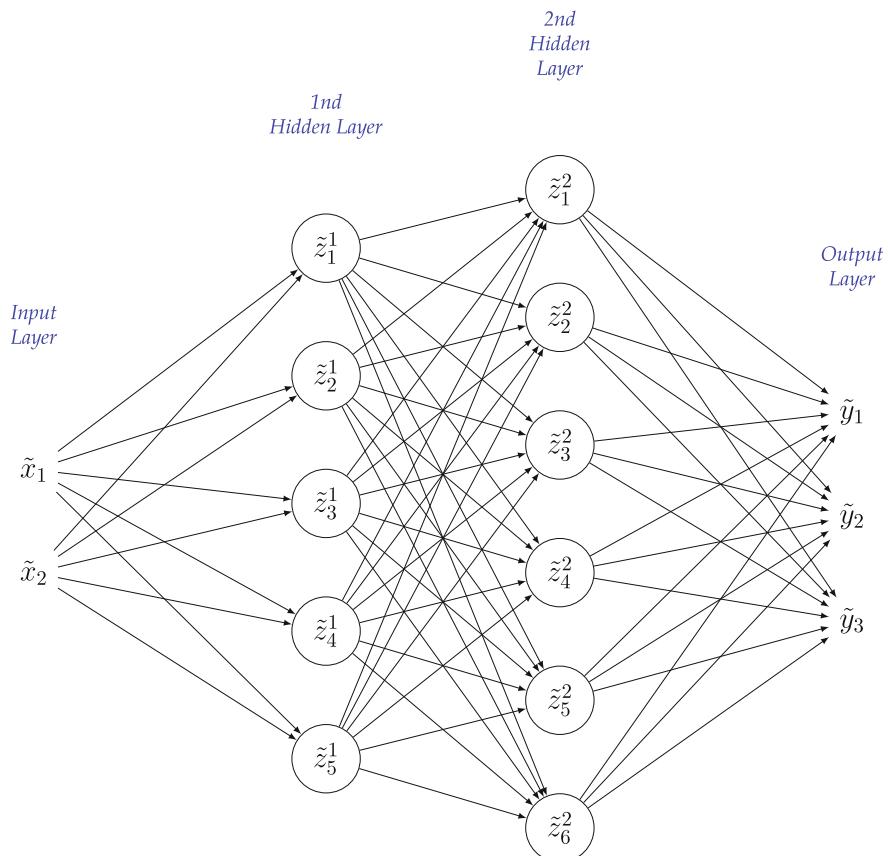


Fig. 1.1 A diagram of a neural network, with 2 inputs, and 3 outputs, and 2 hidden layers of width 5 and 6 neurons respectively.

be the connection (weight). Perhaps a different subclass of neuron would be for each activation function. Do not do this. It will be incredibly slow in almost any programming languages – where objects are always stored by reference, thus requiring a pointer dereferencing for each operation on each neuron. Object orientated is not a good paradigm for this kind of technical computing. Instead program in terms of matrix's and vectors. Potentially using objects for layers, or for the entire network. This allows one to take advantage of efficient matrix algebra routines such as in BLAS and LAPACK.

33 1.2 Function Approximation

34 The Universal Approximation Theorem (UAT) tells us that a neural network can be
35 used to approximate almost any function (Mhaskar et al. 1992; Leshno et al. 1993;
36 Sonoda and Murata 2017). Here a function should be understood in the general sense
37 – classification is a function from observations to classes, regression is a function
38 from observed values to target values. More precisely, for any continuous bounded
39 function, the UAT states that there exists a neural network that can approximate that
40 function, to any given degree of accuracy. Further, it states that the network only needs
41 a single hidden layer. In a less rigorous sense, of approximation, such a continuous
42 bounded function can approximate any function, over a restricted sub-domain of
43 input values.

44 The earliest proofs of this was restricted to bounded activation functions like the
45 sigmoid. More recent work has extended this for unbounded activation functions like
46 ReLU (Sonoda and Murata 2017).

47 A network without any hidden layers is only able to approximate linear functions.
48 It is equivalent to a perceptron, or a linear SVM (Rumelhart et al. 1986; Minsky et al.
49 2017).

50 However, the UAT does not tell us the values for those weight parameters, nor
51 does it promise that any method of training will ever reach them in finite time. More
52 significantly it does not inform as to how wide *sufficiently wide* is, nor if deeper is
53 more efficient than wider.

54 1.3 Network Hyper-Parameters

Hyper-parameters versus Parameters

The weights and biases of the network are called its parameters. They are optimised during training. The other features of the network, including its topology, activation functions, and the training method (which often has its own set of options), are called the hyper-parameters.

55 The key determination in applying neural networks to any problem, is the choice
56 of the hyper-parameters. Including, the number of hidden layers, and their widths.
57 Neural networks can largely be thought of as black-boxes, that can be trained to
58 accomplish tasks given a set of training examples. Beyond recognising how to express
59 the problem, the choice of hyper-parameters is the most important decision when
60 employing a neural network based solution.

61 **1.3.1 Width and Depth**

What happened to Unsupervised Pretraining?

Hinton et al. (2006), Bengio et al. (2007) lead to a new resurgence of interests in neural networks. Pretraining with Deep Belief Networks allowed for deep networks to be trained. It was held for many years that it was necessary to use layer-wise unsupervised pretraining, even when one has only supervised data. Glorot et al. (2011) showed that with ReLU units, a deep network could be trained directly and achieve the same performance. As such unsupervised pretraining is now a more niche technique used primarily when there is an excess of unsupervised data.

62 While the UAT has shown that a single hidden layer is sufficient, it is not necessary
 63 optimal in terms of achieving best performance in practice. For many problems it is
 64 better to have a deeper rather than wider network. This realisation and the techniques
 65 to overcome issues relating to deeper networks lead the current deep learning trend.

66 In the last decade, deep nets have come back into fashion. In brief the reasoning
 67 for this is a combination of better techniques (e.g. ReLU, unsupervised pretraining),
 68 better hardware (e.g. GPUs and Xeon Phi), and more labelled data (e.g. from crowd-
 69 sourcing platforms such as Amazon Mechanical Turk.) This has allowed problems
 70 that were previously considered unsolvable with earlier shallow networks to be solved
 71 with deep networks achieving state of the art results.

72 The choice of the number of hidden layers (depth), and the sizes (width) is a key
 73 parameter in designing a neural network. It is arguably *the* key parameter. It can be
 74 assessed using a hyper-parameter sweep on a validation or development dataset. It
 75 is a particularly relevant parameter for our purposes, as hidden layers normally are
 76 directly linked to our representations of natural language.

77 **1.3.2 Activation Functions**

Notation: Generic Activation Function φ :

Though-out this book, when the activation function is not significant to the problem at hand, we will represent it with the symbol $\varphi(\cdot)$.

Activation functions summary:

identity:

 $(-\infty, \infty)$
range $(-\infty, \infty)$

sigmoid:

range $(0, 1)$

softmax:

range $(0, 1)$

tanh:

range $(-1, 1)$

ReLU:

range $(0, \infty)$

ReLU6:

range $(0, 6)$

The universal approximation theorem places a number of requirements on the choice of activation function. These requirements have been progressively relaxed by works such as Leshno et al. (1993), Sonoda and Murata (2017). This section discusses the activation functions in common use today.

The choice of activation function determines the range of output of a neuron. For the hidden layers these generally perform relatively similarly. Sigmoid, tanh, and ReLU are the most common activation functions used for hidden layers. The range of output of a hidden layer does not normally matter, as it is by definition hidden. More significantly, the final activation function or the output function does always matter. The output function should be chosen for the task at hand. As will be discussed in Chap. 3 this also has significance when designing gated neural network structures.

Beyond determining the range of the output, its derivative also determines the behaviour during training via gradient descent based training methods. This is significant for activation functions like ReLU and ReLU6 discussed in the following sections.

The activation functions, other than the softmax layer, are applied element-wise to vector inputs. As such the scalar definitions are given here. We write $y = \varphi(z)$, for y being the output, and z being the weighted input from the layer below. To represent the whole layer including weights and biases we write:

$$\tilde{y} = \varphi(z) = \varphi(W\tilde{x} + \tilde{b}) \quad (1.1)$$

1.3.2.1 Identity (Affine/Linear)

The most basic activation function is the identity. To simply output the weighted sum of the inputs, without applying a non-linearity. It is commonly used as an output function for regression tasks, as it imposes no (additional) bounds.

$$\tilde{y} = \tilde{z} \quad (1.2)$$

Often the use of the identity activation function will be described in terms of the whole layer, including the weight and bias. The layer is often called an affine layer as it corresponds to an affine transformation of its input. When the bias is fixed at zero, this is a linear transformation of the input.

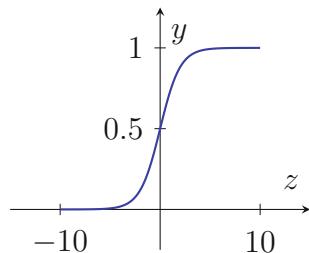
$$\tilde{y} = W\tilde{x} + \tilde{b} \quad (1.3)$$

In this equation W and \tilde{b} are a trained weight matrix and bias vector respectively, and \tilde{x} is the output of the layer below.

Using an affine output layer on-top of the hidden layers allows the learned scaling and shifting of any of the other activation functions. However, when this is not needed it increases the difficulty of training for no benefit.

Identity functions cannot be used as hidden layer activation functions. A stack of affine layers simplifies to be equivalent to a single affine layer, which means that

Fig. 1.2 The sigmoid function



the network can only approximate linear functions. A non-linearity is required, as is provided by most activation functions.

1.3.2.2 Sigmoid

Sigmoid is the classic neural network activation function. It is shown in Fig. 1.2.

AQ1

$$y = \sigma(z) = \frac{1}{1 + \exp(-z)} \quad \text{LW: add brackets} \quad (1.4)$$

Sigmoid outputs values between zero and one. If the network has a single output then this is a fuzzy boolean. It can be used to classify into two categories. It is sometimes referred to as the logistic output function as it forms the basis of logistic regression. (Conversely, sometimes the word sigmoid is used to refer to all functions of such an S shape.)

Note that this function has the property that, $\sigma(-z) = 1 - \sigma(z)$. This means negating the input to the final sigmoid in a binary classification network, is the same as switching which class is considered as true.

Multinomial logistic regression

A neural network classifier with a softmax output layer is sometimes called a multinomial logistic regression network (especially if there are no hidden layers), or even just a logistic regression network. This can lead to some confusion with a sigmoid outputting network. However, the similarity can be understood by considering a softmax output layer with 2 classes. For $\tilde{z} = [\tilde{z}_1, \tilde{z}_2]$ defining $v = \tilde{z}_1 - \tilde{z}_2$

A (binary) logistic classification using the same inputs can be written: $P(Y = 1) = \sigma(v) = \frac{1}{1 + \exp(\tilde{z}_2 - \tilde{z}_1)}$. The 2 class softmax formulation is:

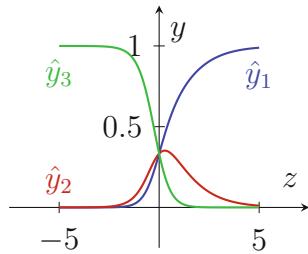
$$\begin{aligned} & \text{smax}(\tilde{z}) \\ &= \left[\frac{\exp \tilde{z}_1}{\exp \tilde{z}_1 + \exp \tilde{z}_2}, \frac{\exp \tilde{z}_2}{\exp \tilde{z}_1 + \exp \tilde{z}_2} \right] \\ &= \left[\frac{1}{1 + \exp(\tilde{z}_2 - \tilde{z}_1)}, \frac{1}{\exp(\tilde{z}_1 - \tilde{z}_2) + 1} \right] \\ &= [\sigma(v), \sigma(-v)] \\ &= [\sigma(v), 1 - \sigma(v)] \\ &= [P(Y = 1), P(Y = 2)] \end{aligned}$$

LW: break line to avoid splitting equation across lines

LW: Indent all = lines under smax



Fig. 1.3 An example of a softmax activation in use for with 3 classes for $z \mapsto \text{smax}([z; \frac{z}{10}; -2z])$. Notice at all points the 3 outputs sum to 1



1.3.2.3 Softmax

The standard output activation function for multiclass classification is softmax. Formally, training a so-called softmax classifier is regression to a class probability, rather than a true classifier as it does not return the class but rather a prediction of each class's likelihood. We will use the terms interchangeably.

Defining the layer \tilde{y} in terms of its elements, recall that by our notation convention \tilde{y}_i is the i th element of the vector $\tilde{y} = [\tilde{y}_1, \tilde{y}_2, \dots, \tilde{y}_n]$:

$$\tilde{y}_i = (\text{smax}(\tilde{z}))_i = \frac{\exp \tilde{z}_i}{\sum_{j=1}^N \exp \tilde{z}_j} \quad (1.5)$$

For example as shown in Fig. 1.3.

\tilde{y} is a vector of length n , where n the number of classes in the classification.

These elements have values between zero and one, and sum to one. They thus define a discrete probability mass function (Fig. 1.3).

When used for classification \tilde{y}_i gives the probability of being in class i .

$$P(Y = i \mid \tilde{z}) = (\text{smax}(\tilde{z}))_i = \tilde{y}_i \quad (1.6)$$

Softmax is a *soft*, i.e. differentiable, version of what could be called *hard-max*. In this conceptual hard-max function, the largest output is set to one, and the others set to zero. It is non-differentiable and thus not suitable for use in a neural network. In softmax the larger values are proportionally increased to be closest to one, and the smaller to be closest to zero.

Further consideration of the softmax is given in Sect. 2.5.1.1.

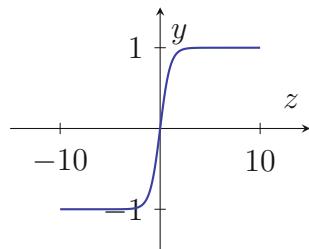
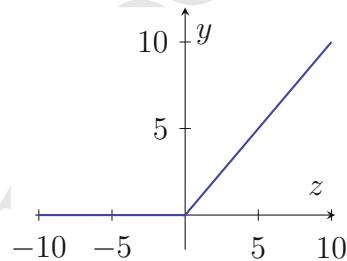
1.3.2.4 Tanh

The hyperbolic tangent function is a scaled and shifted sigmoid function. Such that the outputs are restricted to be between -1 and $+1$ (Fig. 1.4).

It is shown in

$$y = \tanh(z) = \frac{\exp(2z) - 1}{\exp(2z) + 1} = 2\sigma(2z) - 1 \quad (1.7)$$



Fig. 1.4 The tanh function**Fig. 1.5** The ReLU function

1.3.2.5 ReLU

The Rectified Linear Unit (ReLU) is a more recently popularized activation function (Dahl et al. 2013). It comes from earlier neuroscience work (Hahnloser 1988).

Fig. 1.5.

It is shown in

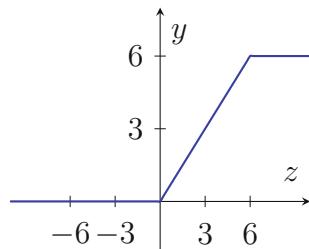
$$y = \text{ReLU}(z) = \max(0, z) = \begin{cases} z & 0 \leq z \\ 0 & z < 0 \end{cases} \quad (1.8)$$

Values are restricted to be at non-negative. As this function has derivative zero for $z < 0$, once a unit is turned off, it is not turned back-on. During gradient descent (see Sect. 1.4.0.1) the derivative is used to modify the weights, as it is zero it the weights will never change once turned off. This is commonly called the neuron dying. Further to this, using (zero mean) random initialization ensures that a significant portion of all neurons will initially be dead. This results in sparse connections between the layers. This has been found to be a good thing (Glorot et al. 2011). ReLU is very commonly used as a hidden layer activation function for deep networks. It also helps with the gradient vanishing and gradient exploding problems that occur in deep networks (Glorot et al. 2011).

1.3.2.6 RELU6

Closely related to ReLU is ReLU6 (Krizhevsky et al. 2010). This is another linear unit that saturates at 0, but also at 6 (Fig. 1.6).

Fig. 1.6 The RELU6 function



170

$$y = \text{ReLU}_6(z) = \max(0, \min(6, z)) = \begin{cases} 6 & 6 < z \\ z & 0 \leq z \leq 6 \\ 0 & z < 0 \end{cases} \quad (1.9)$$

Why are we bringing up ReLU6?

ReLU6 is one of the more obscure activation functions. As such it may seem a bizarre inclusion in this shorter introduction. However, we have found it surprisingly successful as an activation function, e.g. in the auto-encoder example which follows in Sect. 1.5.2. It also nicely illustrates the flexibility and indeed arbitrariness in how activation functions can be defined.

171
172
173
174
175

This has similar advantages to ReLU. However, as well as units being able to die to zero, it can also die to positive. The bounding at 6, rather than any other number is rather arbitrary – particularly given scaling with the intervening affine layers. The important point is that it is bound both above and below. This greatly simplifies any proofs of theoretical capabilities of various networks.

176
177
178
179

1.3.2.7 Other Activation Functions

There are numerous other activation functions. Most are of interest for use in hidden layers such as Leaky RELU, Softplus, ELU and many others. For specialized tasks, other functions might be used, such as atan2 (White and Lydon 2016).

180

1.4 Training

181
182

The process of training the network is the process of solving a very high dimensional, nonlinear, non-convex global optimization problem.

Evaluation function versus Loss function

We must distinguish between two types of *error function*. The **evaluation function** is the metric we are truly trying to improve, this could be accuracy for classification, BLEU score for translation, F1 score for retrieval etc. It is applied to the whole system (which may be greater than just the neural network), and the system may be evaluated in different ways using different evaluations. As the evaluation is often not differentiable, a proxy for it which we call the **loss function** is used in training. For example, squared error for regression, or cross-entropy for classification. The loss function is applied to the output of the network during training to calculate the error between the network output for a given input, and the target output from the training example. The loss function is designed such that minimising the loss function also (ideally) results in the true evaluation function being optimised.

183 Finding the globally optimal solution to such a function is itself a very difficult
 184 problem. Nonlinear and non-convex problems can have local minima that are not
 185 global optima. This means they cannot guarantee to reach a global optima by gradient
 186 descent. However, this does not pose an issue for most neural networks as finding
 187 the true global optima is not required, merely finding a *good enough* set of weights
 188 to get a *good enough* result.

189 To train the network one must find the values for the networks weight and bias
 190 parameters, such that the total loss function is minimized over the training set.

191 To define the training problems as an optimisation problem one first considers the
 192 loss function for a single training case. For a target output y and an actual output \hat{y} ,
 193 (considering the scalar case) a loss function is defined $loss(y, \hat{y})$. For example the
 194 squared error loss (sometimes called L2 loss, though that risks confusion with L2
 195 weight penalty for regularisation) used in regression is defined by

$$196 \quad SE(y, \hat{y}) = (y - \hat{y})^2 \quad (1.10)$$

197 The cross-entropy loss used in binary classification is defined by

$$198 \quad CE(y, \hat{y}) = -((1 - y) \log(1 - \hat{y}) + y \log(\hat{y})) \quad (1.11)$$

199 The choice of loss function depends on the purpose of the network.

200 When network (represented by the function f) is composed into the loss function
 201 the per training case loss is given by $loss(\tilde{y}, \tilde{f}(\tilde{x}, \theta))$. $\tilde{f}(\tilde{x}, \theta)$ is the function that
 202 executes some neural network, with weight and bias parameters all grouped into θ ,
 203 upon an input \tilde{x} where the target output is \tilde{y} .

204 The total loss ($Loss$) is defined by taking the sum over the whole training set (\mathcal{X})
 205 of input output pairs.

$$206 \quad Loss(\theta, \mathcal{X}) = \sum_{\forall (\tilde{x}, \tilde{y}) \in \mathcal{X}} loss(\tilde{y}, \tilde{f}(\tilde{x}, \theta)) \quad (1.12)$$

207 Equation (1.12) nothing but a nonlinear function to be minimised by adjusting the
 208 values of the weights and biases in θ .

209 This can be given to off-the-shelf unconstrained nonlinear optimisation algorithms
 210 (Ngiam et al. 2011) such as L-BFGS (Nocedal and Jorge 1980). Alternatively, and
 211 more commonly, the loss and updated can be processed iteratively on sub-sets, called
 212 minibatches, of the training data at a time. The extreme case of this is to update for
 213 every single training example – this is called online training. Updating on minibatches
 214 is generally preferred over updating per example as it results in less small “noisy”
 215 weight fluctuations which may prevent reaching local minima. It is preferred over
 216 full-batch (on the whole training set) as the more frequent updates allow for faster
 217 learning (LeCun et al. 2012). Minibatch optimisation is often performed using algo-
 218 rithms specifically targeted at machine learning applications such as Adam (Kingma
 219 and Ba 2014). In either case, almost all optimizers used for neural networks rely on
 220 gradient descent.

AdaGrad, AdaDelta, and Adam

AdaGrad (Duchi et al. 2011), AdaDelta (Zeiler 2012) and Adam (Kingma and Ba 2014) are effectively successive versions of an optimisation algorithm, specifically targeted at the machine learning case. These algorithms dynamically adapt the learning rate during training. There are several other algorithms in the family, for example RMSprop. In recent works, Adam is the most commonly used optimiser. Dauphin et al. (2015) presents another algorithm from this family, and considers how the adaptive learning rate is related to the second order derivative, which is the fundamental innovation in more traditional quasi-Newtonian methods (over first-order methods like plain gradient descent.).

221 1.4.1 Gradient Descent and Back-Propagation

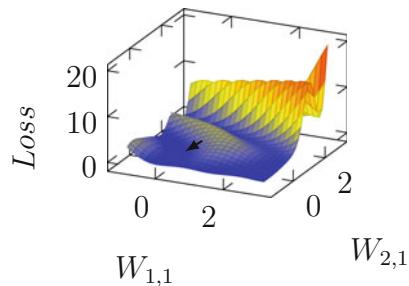
Back-propagation versus Gradient Descent?

Formally speaking, back-propagation finds the gradients, and gradient descent uses those gradients to update the network parameters. However, the terms are often used interchangeably.

222 Gradient descent is the basis of most commonly used optimisers in machine learning.
 223 In gradient descent based methods the derivatives of the loss function relative to the
 224 parameters are used to update those parameters. The parameters in this case are
 225 the network weights and biases. To calculate the gradients the back-propagation
 226 algorithm is used.

227 Back-propagation is simply a method of applying the well-known chain-rule of
 228 calculus to the neural network loss functions (Rumelhart et al. 1986). It allows one to
 229 calculate: $\frac{\partial \text{Loss}}{\partial W_{i,j}}$ for any weight (or bias) parameter $W_{i,j}$ in the network. These partial

Fig. 1.7 A visualisation of a possible error (loss) surface. The arrow indicates a possible step from one set of parameters (weights) down to another set of parameters with lower loss, via gradient descent.



derivatives can always be found as the overall loss function, including the network, is differentiable¹ (when using typical activation functions and loss functions).

Using these gradients the network parameters can thus be updated. Updating the parameters to a new value in the direction of the gradient decreases the loss (for the right step size). This means for each $W_{i,j} \in \theta$ updating its value according to the local gradient.

$$W_{i,j} \leftarrow W_{i,j} - \alpha \frac{\partial Loss}{\partial W_{i,j}} \quad (1.13)$$

where α is an update step size, commonly called the learning rate. The more advanced methods discussed earlier, such as Adam, and L-BFGS, are extensions along this core principle.

If one considers a plot with the loss given on the vertical axis, and the values of the parameters as describing a position on the horizontal axes, then gradient descent is moving that the parameter-point downward on that surface based on local estimate of the slope. Such a plot is shown in Fig. 1.7. Most real networks have far more than 2 parameters, of-course, however the process of gradient decent remains the same.

These examples are available online

For a practical introduction to the networks discussed here, worked examples of these network are available in the accompanying blog post at <http://white.ucc.asn.au/NNforNLPBook/NNexamples>. They are written in the Julia programming language, making use of the TensorFlow.jl package.

¹Technically most networks are differentiable almost everywhere. For some functions like ReLU they have some single points of where the derivative is not defined. But the derivatives here can be forced to a known value.

245 1.5 Some Examples of Common Neural Network Architectures

246 We will discuss more linguistically relevant neural networks in the following chapters.
 247 However, to introduce the topic, we will give some basic examples.

FashionMNIST

Is a new image classification benchmarking dataset (Xiao et al. 2017). It is a task to classify images of shirts, shoes and bags etc. It is an exact drop-in replacement for MNIST. It was created to allow the same benchmarking scripts to be used; while using a different data set. This combats extremely high use of MNIST leading to implicit test set information leakage in the design of ML systems (where systems are base on systems which perform well on MNIST due to more of such systems being published.); as well as being a more difficult task, thus allowing better discrimination between systems.

248 1.5.1 Classifier

249 A classifier on the MNIST challenge is one of the most common introductory net-
 250 works. The MNIST dataset is a collection of images of hand written digits, which
 251 much be classified as to which digit they are.

252 A basic network to complete this task is is shown in Fig. 1.8.

253 The network is defined by the equations:

$$254 \quad \tilde{z} = \varphi(W^1 x + \tilde{b}_1) \quad (1.14)$$

$$255 \quad \hat{y} = \text{smax}(W^2 z + \tilde{b}_2) \quad (1.15)$$

$$256 \quad P(Y = i | X = \tilde{x}) = \hat{y}_i \quad (1.16)$$

$$257 \quad 258 \quad = \left[\text{smax}(W^2 \varphi(W^1 \tilde{x} + \tilde{b}^1) + \tilde{b}^2) \right]_i \quad (1.17)$$

259 for X the input variable of grey-scale pixels intensities, and Y the output as a class.
 260 The output is represented with 1 hot encoding, where the index corresponding to the
 261 class is 1, and the others zero.

This simple network performs very well

We note that this simple network can without further enhancement exceed the early published results for the MNIST challenge, using standard un-augmented neural networks, simply by using a very wide hidden-layer, that was not feasible at the time of those benchmarks. It is still outperformed by convolution neural networks and other better architectures.

262 To go into details about each layer: The input is given by the grey scale pixel
 263 intensities in the original 28×28 image. This is flattened giving a 784 element sized

Probabilities of each of the classes y_i :

$$P(Y=i|X=\tilde{x})$$

LW: this image seems to be too large and should be scaled down so the text is similar in size to the main text

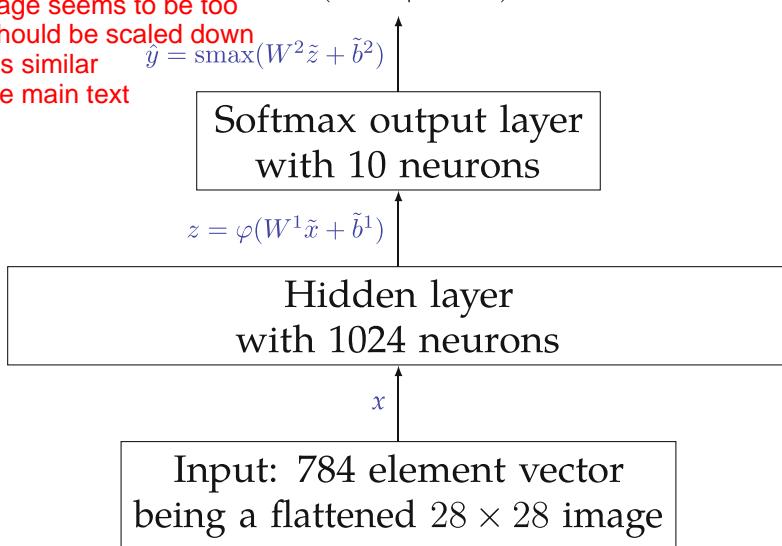


Fig. 1.8 The structure of a simple classifier network for MNIST.

vector (\tilde{x}). The first weight matrix projects that up onto a hidden layer of width 1024. Thus W^1 is a 1024×784 matrix. The bias \tilde{b}^1 is for that hidden layer, so is a vector of length 1024. The hidden layer's actual activation values (\tilde{z}) for any input can be considered as being the values taken by 1024 different latent variables describing that input. These are chosen from a continuous abstract space of variables (via training) to effectively discriminate the correct class in the next layer (the output layer).

This is a shallow network, with only one hidden layer. A deep network would have more hidden layers, for additional latent variables describing relations. For the output layer we must project down to 10 dimensions, as there are 10 classes to choose from: the numbers from 0 to 9. thus W^2 is a 10×1024 matrix, and the bias \tilde{b}^2 is a 10 element vector. Using the softmax layer here ensures that output \hat{y} is a valid probability mass function (nonnegative and summing to one).

AQ2

1.5.1.1 Softmax and Bayes' Theorem

As a digression, it is worth considering the similarity between a network with a softmax output layer, and the application of Bayes' Theorem. This will become important for understanding output embeddings, and hierarchical softmax in the future chapters.

The conditional probability of a classification given the value of some observed variable is defined by $P(Y = y | Z = \tilde{z})$. For Y being the class taking value i ; and Z the variable being conditioned upon, taking value \tilde{z} . Here \tilde{z} is some feature vector,



this could be the output of a hidden layer below, or it could be a direct input. Using softmax the estimated probability is given by

$$P(Y = i \mid Z = \tilde{z}) = \text{smax}(\tilde{z})_i \quad (1.18)$$

$$= \frac{\exp(W\tilde{z} + \tilde{b})_i}{\sum_{\forall j} \exp(W\tilde{z} + \tilde{b})_j} \quad (1.19)$$

$$= \frac{\exp(W\tilde{z})_i \exp(\tilde{b}_i)}{\sum_{\forall j} \exp(W\tilde{z})_j \exp(\tilde{b}_j)} \quad (1.20)$$

LW: Font on exp should be math-roman, not typewriter

One can see that the bias term $\exp \tilde{b}_i$ does not depend on the value of z . The bias term is analogous to the prior probability. Literally, it is the bias towards each element (i.e. each value Y could take) without observing the condition. We will define an unnormalised probability-like scoring function R ; and say $R(Y = i) = \exp \tilde{b}_i$, representing the marginal score contribution towards class i from the bias.

The other component is $\exp(W\tilde{z})_i$. By considering this for each index i (class value) that Y might take then we have

$$(W\tilde{z})_i = \sum_{\forall j} W_{i,j} \tilde{z}_j = W_{i,:}\tilde{z} \quad (1.21)$$

Given one is considering the case for a particular class $Y = i$, then it can be seen that the row vector $W_{i,:}$ as a weighting map for features possessed by \tilde{z} , i.e. for input (Z) feature j , $W_{i,j}$ determines the relative weighting for j th feature versus the other features of Z , when the class is i . When trained, the weight values will be such that it scores how likely the features are to occur for a given output class.

We can say that our scoring function R , has a likelihood term:

$$R(Z = \tilde{z} \mid Y = i) = \exp(W\tilde{z})_i \quad (1.22)$$

We will return to these row vectors $W_{i,:}$ when we discuss as output embeddings in Chap. 4. $W_{i,:}$ must characterise the output class i , since different classes would assign different importance to different features in the Z . Furthermore, to some degree similar classes, i.e. classes triggered by similar features, would thus have more similar weights.

We can combine the unnormalised score factors to reformulate the original statement:

$$P(Y = i \mid Z = \tilde{z}) = \frac{R(Z = \tilde{z} \mid Y = i) R(Y = i)}{\sum_{\forall j} R(Z = \tilde{z} \mid Y = j) R(Y = j)} \quad (1.23)$$

313 Contrast this to Bayes' Theorem:

$$\text{314} \quad P(Y = i | Z = \tilde{z}) = \frac{P(Z = \tilde{z} | Y = i) P(Y = i)}{P(Z = \tilde{z})} \quad (1.24)$$

$$\text{315} \quad = \frac{P(Z = \tilde{z} | Y = i) P(Y = i)}{\sum_{\forall j} P(Z = \tilde{z} | Y = j) P(Y = j)} \quad (1.25)$$

316

Marginal Denominator in Bayes Theorem

In Eq. (1.25), notice that we replace $P(Z = \tilde{z})$. We do this using the rule for finding marginal probabilities, from conditional probabilities with mutually exclusive and exhaustive condition values. This is always possible when working with classification.

317 The similarities can be seen. The bias effectively determines a prior probability-like term. The weights defines the posterior probability: that is the chance of a 318 particular class having a particular set of features. The rows of the weight matrix 319 mark how important each hidden feature is to each class. We can consider that each 320 row of the weight matrix is itself a representation of the class, as characterised by 321 the importance of those latent features. In the next chapter we will consider this as 322 an output embedding (Fig. 1.9).

323 A more traditional way of finding a representation of an input (rather than an 324 output class) is to use an autoencoder.

326 **1.5.2 Bottlenecking Autoencoder**

327 An autoencoder is a tool primarily used for finding a representation of their input. 328 There are many varieties of autoencoder based on neural network related techniques, 329 including the works of Hinton (2002), Hinton and Salakhutdinov (2006), Hinton et al. 330 (2006), Vincent et al. (2010), Chen et al. (2012), Kingma and Welling (2014). This 331 itself is a whole sub-field of machine learning. Here we look at a simple bottleneck- 332 autoencoder (Bourlard et al. 1988; Japkowicz et al. 2000). It has been used in 333 a variety of tasks to attempt to find an optimal representation for an input e.g. as in 334 Usui et al. (1992). An autoencoder is a neural network tasked with outputting its 335 input. This seemingly is a pointless task – one already has the perfect reproduction 336 of the input, in the input itself. However, the true use of an autoencoder is to extract 337 the output of one of the intermediary layers. We call the intermediary layer the code 338 layer. To force this coded representation to have useful interesting properties, and 339 to prevent the network from simply learning the identity function, all autoencoder 340 include one or more features that increases the difficulty. In the case of the bottlenecking 341 autoencoder this feature is the bottleneck. The code layer is much narrower 342 than the input layer. This forces the network to effectively learn to compress the data 343 – performing dimensionality reduction.

LW: Image should be scaled to fit with normal orientation on page.



Fig. 1.9 A sampling of MNIST images from the test-set, arranged according to the values of 2 neurons on the encoding layer.

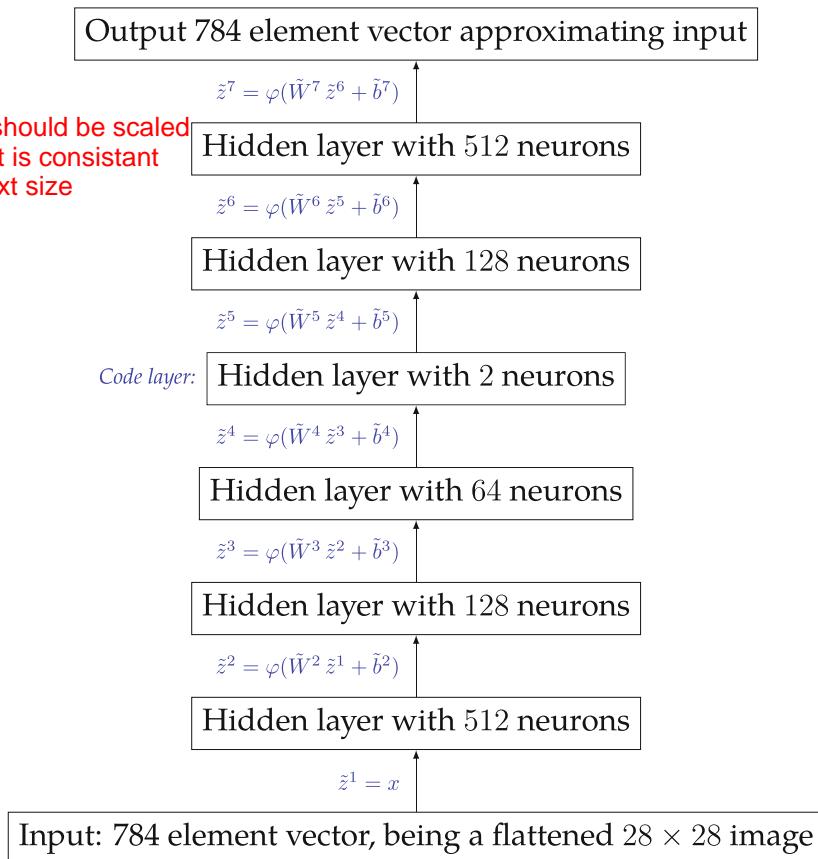


Fig. 1.10 A structure for a bottle necking autoencoder for use on MNIST.

In this particular example we are looking at an auto-encoder for the MNIST images discussed earlier. The original images are 28×28 pixels, i.e. 784 dimensional. We compress it down to just 2 dimensions using the network shown in Fig. 1.10.

In this particular network φ is a leaky RELU6 unit. As shown in Fig. 1.11

$$\varphi(z) = \begin{cases} 0.01z + 6 & 6 < z \\ 1.01z & 0 \leq z \leq 6 \\ 0.01z & z < 0 \end{cases} \quad (1.26)$$

The reason for this is that a sigmoidal unit does not perform well in a deep network because of the gradient-vanishing/exploding effects. As a result, only the bias information of the final layer mattered, effectively making all outputs just the average of all inputs. The normal solution to this in deep networks is ReLU or ReLU6 units (Fig. 1.11).



Fig. 1.11 The Leaky ReLU6 function. The leak level on this plot is exaggerated for purposes of visualisation.

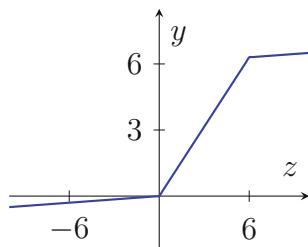
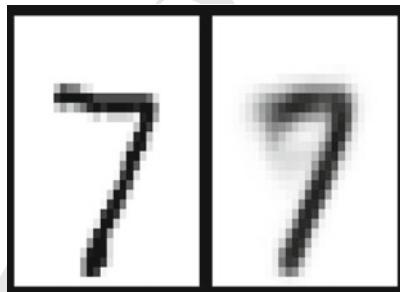


Fig. 1.12 Recreation of an input from the MNIST test set. Input is on the left, output is on the right.



However, a ReLU, or ReLU6 unit is not ideal either, as these units turn-off, and cannot turn back on again. If one of the only 2 neurons in the code layer turn-off then they cannot turn back on again – thus forcing it down to just one neuron, or none if that neuron also dies. In trialling this network structure, this was found to occur quite very often.

The solution was to add a “leak” to the unit. A small constant gradient even in the saturated position. Making such a neuron possible, even if difficult, to turn back on once turned off. With this change the network always produces quality results. An example of the recreation of an input from the test set can be seen in Fig. 1.12. It can be seen that the blur suggests that the input is a 7 or 9 like shape – that level of information survived being squeezed through the bottleneck.

More significantly, a sampling of the code layer is visually shown in Fig. 1.9. In this image, the positions on the X and Y axis for each input image is given by the values from the code layer for the representation of that image. It can be seen that the code is capturing key information about the appearance. The 0s appear near to the similarly rounded 6s, the 8s near the 3s, etc. In particular the 1s are arrayed according to their slope. The autoencoder has captured very useful information about the inputs, that would be hard to define with any hand-written feature extractor.

It is this property of networks, to implicitly discover and define the most important latent variables and relationships to the task at hand that also makes it valuable for natural language processing.

References

- Bengio, Yoshua, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. 2007. Greedlayer-wise training of deep networks. In *Advances in neural information processing systems*, 153–160.
- Bourlard, Herve, and Yves Kamp. 1988. Auto-association by multilayer perceptrons and singular value decomposition. *Biological Cybernetics* 59 (4): 291–294.
- Chen, Minmin, Zhixiang Xu, Kilian Weinberger, and Fei Sha. 2012. Marginalized denoising autoencoders for domain adaptation. In *Proceedings of the 29th international conference on machine learning (ICML-12)*, eds. John Langford and Joelle Pineau, 767–774. Edinburgh: Omnipress. ISBN: 978-1-4503-1285-1.
- Dahl, George E., Tara N. Sainath, and Geoffrey E. Hinton. 2013. Improving deep neural networks for LVCSR using rectified linear units and dropout. In *2013 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 8609–8613. IEEE. <https://doi.org/10.1109/ICASSP.2013.6639346>.
- Dauphin, Yann N., Harm de Vries, Junyoung Chung, and Yoshua Bengio. 2015. RMSProp and equilibrated adaptive learning rates for non-convex optimization. In *CoRR*. arXiv:abs/1502.04390.
- Duchi, John, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research* 12: 2121–2159.
- Glorot, Xavier, Antoine Bordes, and Yoshua Bengio. 2011. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 315–323.
- Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning*. MIT Press. <http://www.deeplearningbook.org>
- Hahnloser, Richard L.T. 1998. On the piecewise analysis of networks of linear threshold neurons. *Neural Networks* 11 (4): 691–697.
- Hebb, D.O. 1949. *The organization of behavior: A neuropsychological theory*. A Wiley book in clinical psychology. New York: Wiley.
- Hinton, Geoffrey E. 2002. Training products of experts by minimizing contrastive divergence. *Neural Computation* 14 (8): 1771–1800.
- Hinton, Geoffrey E., Simon Osindero, and Yee-Whye Teh. 2006. A fast learning algorithm for deep belief nets. *Neural Computation* 18 (7): 1527–1554
- Hinton, Geoffrey E. and Ruslan R. Salakhutdinov. 2006. Reducing the dimensionality of data with neural networks. *Science* 313 (5786): 504–507.
- Japkowicz, Nathalie, Stephen Jose Hanson, and Mark A. Gluck. 2000. Nonlinear auto association is not equivalent to PCA. *Neural Computation* 12 (3): 531–545.
- Kingma, D. P. and M. Welling. 2014. Auto-encoding variational bayes. In *The international conference on learning representations (ICLR)*. arXiv:1312.6114 [stat.ML].
- Kingma, Diederik and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv:1412.6980.
- Krizhevsky, Alex and G. Hinton. 2010. Convolutional deep belief networks on CIFAR-10. *Unpublished Manuscript* 40.
- LeCun, Yann A., Leon Bottou, Genevieve B. Orr, and Klaus-Robert Muller. 2012. Efficient backprop. *Neural networks: Tricks of the trade*, 9–48. Berlin: Springer.
- Leshno, Moshe, Vladimir Ya. Lin, Allan Pinkus, and Shimon Schocken. 1993. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks* 6 (6): 861–867.
- Mhaskar, Hrushikesh N., and Charles A. Micchelli. 1992. Approximation by superposition of sigmoidal and radial basis functions. *Advances in Applied Mathematics* 13 (3): 350–373.
- Minsky, Marvin, Seymour A. Papert, and Leon Bottou. 2017. *Perceptrons: An introduction to computational geometry*. Cambridge: MIT Press.

- 425 Ngiam, Jiquan, Adam Coates, Ahbik Lahiri, Bobby Prochnow, Quoc V. Le, and Andrew Y. Ng. 2011.
426 On optimization methods for deep learning. In *Proceedings of the 28th international conference
427 on machine learning (ICML-11)*, 265–272.
- 428 Nielsen, Michael A. 2015. *Neural networks and deep learning*. Determination Press.
- 429 Nocedal, Jorge. 1980. Updating quasi-Newton matrices with limited storage. *Mathematics of Com-
430 putation* 35 (151): 773–782.
- 431 Rumelhart, David E., Geoffrey E. Hintont, and Ronald J. Williams. 1986. Learning representations
432 by back-propagating errors. *Nature* 323: 9.
- 433 Sonoda, Sho and Noboru Murata. 2017. Neural network with unbounded activation functions is
434 universal approximator. *Applied and Computational Harmonic Analysis* 43 (2), 233–268. ISSN:
435 1063-5203, <https://doi.org/10.1016/j.acha.2015.12.005>.
- 436 Usui, Shiro, Shigeki Nakaochi, and Masae Nakano. 1992. Reconstruction of Munsell color space by
437 a five-layer neural network. *Journal of the Optical Society of America A* 9 (4): 516–520. <https://doi.org/10.1364/JOSAA.9.000516>.
- 438 Vincent, Pascal, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre- Antoine Manzagol.
439 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a
440 local denoising criterion. *Journal of Machine Learning Research* 11: 3371–3408.
- 441 White, Lyndon. 2016. Encoding angle data for neural networks. Cross validated stack exchange.
442 <https://math.stackexchange.com/q/2369786>.
- 443 Xiao, Han, Kashif Rasul, and Roland Vollgraf. 2017. Fashion-MNIST: A novel image dataset
444 for benchmarking machine learning algorithms. [arXiv:1708.07747](https://arxiv.org/abs/1708.07747) [cs.LG], <https://github.com/zalandoresearch/fashionmnist>.
- 445 Zeiler, Matthew D. 2012. ADADELTA: An adaptive learning rate method. In *CoRR*.
446 [arXiv:abs/1212.5701](https://arxiv.org/abs/1212.5701).

Author Queries

Chapter 1

Query Refs.	Details Required	Author's response
AQ1	Please check and confirm if the inserted citation of Figs. 1.2–1.6 and 1.11 are correct. If not, please suggest an alternate citation. Please note that figures should be cited sequentially in the text.	
AQ2	Please note that mismatch have been found between pdf and tex regarding the parenthesis in equations hence we have followed tex source. Kindly check and confirm.	

LW:

AQ1: I have suggested alternate citation inline.

AQ2: I am not sure exactly which equation you mean, but they seem correct to me. However:

It seems that the brackets have been removed from most of the exp in the math
I suspect my override of \exp has been lost somewhere.
Which does make a lot of the math wrong (or at least unclear).
I've tried to mark it when I spotted it inline

Note that the preamble contained:

```
```\let\oldexp\exp  
\renewcommand{\exp}[1]{\oldexp \left(#1 \right)}
```

Also

I have noticed that the final period seems to have been removed from all the captions. I have marked it where i spotted it is missing, but I believe that it is likley an issue in the processing system.

# Metadata of the chapter that will be visualized in SpringerLink

Book Title	Neural Representations of Natural Language	
Series Title		
Chapter Title	Recurrent Neural Networks for Sequential Processing	
Copyright Year	2019	
Copyright HolderName	Springer Nature Singapore Pte Ltd.	
Corresponding Author	Family Name Particle Given Name Prefix Suffix Role Division Organization Address Email	<b>White</b> <b>Lyndon</b> <b>Togneri</b> <b>Roberto</b> <b>Liu</b> <b>Wei</b> <b>Bennamoun</b> <b>Mohammed</b>
Author	Family Name Particle Given Name Prefix Suffix Role Division Organization Address Email	School of Electrical, Electronics and Computer Engineering The University of Western Australia Perth, WA, Australia LYNDON.WHITE@RESEARCH.UWA.EDU.AU
Author	Family Name Particle Given Name Prefix Suffix Role Division Organization Address Email	School of Electrical, Electronics and Computer Engineering The University of Western Australia Perth, WA, Australia
Author	Family Name Particle Given Name Prefix Suffix Role Division Organization Address Email	School of Computer Science and Software Engineering The University of Western Australia Perth, WA, Australia
Author	Family Name Particle Given Name Prefix Suffix	

Role  
Division School of Computer Science and Software Engineering  
Organization The University of Western Australia  
Address Perth, WA, Australia  
Email

---

Abstract	This chapter continues from the general introduction to neural networks, to a focus on recurrent networks. The recurrent neural network is the most popular neural network approach for working with sequences of dynamic size. As with the prior chapter, readers familiar with RNNs can reasonably skip this. Note that this chapter does not pertain specifically to NLP. However, as NLP tasks are almost always sequential in nature, RNNs are fundamental to many NLP systems
----------	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------



# Recurrent Neural Networks for Sequential Processing

2

LW: Where is the chapter abstract gone?

*I've the RNN with and works, but the computed with program of the RNN with and the computed of the RNN with with and the code — The output of an RNN created by Andrej Karpathy (2015), trained on an article on the use of RNNs for generating text (it works poorly due to the very low amount of training data)*  
<http://karpathy.github.io/2015/05/21/rnneffectiveness/>

LW:  
This section should not be italic. it is the source of the quotation.

- 1 This chapter continues from the general introduction to neural networks, to a focus on  
2 recurrent networks. The recurrent neural network is the most popular neural network  
3 approach for working with sequences of dynamic size. As with the prior chapter,  
4 readers familiar with RNNs can reasonably skip this. Note that this chapter does not  
5 pertain specifically to NLP. However, as NLP tasks are almost always sequential in  
6 nature, RNNs are fundamental to many NLP systems.

## Works generated by RNNs

LW: spaces missing

Generating text using RNNs is very hip, and pretty rad. <https://arstechnica.com/?p=896589>, <https://github.com/zackthout/got-book-6>, <http://aiweirdness.com/post/168770625987/>. The results are often surprisingly cognisant, and almost always entertaining. It should be noted however, that this is in-principle little different to sampling Markov chains from probabilistic language models. Even the longest of longer short term memory systems still do not truly have the memory to actually sensibly write prose; which would require memory of the outputs from multiple paragraphs (or chapters) ago.



## 7 2.1 Recurrent Neural Networks

### Time-Step

RNNs are normally described in terms of a time-step. This is the advancement of the system such that the previous output state, is now the input old-state. This does not have to be literal time – indeed it cannot be, as it is discrete. In most NLP applications it is time analogous: words in the order they are said. In other NLP applications it might actually be words in the reverse order to that in which they are normally said. In other machine learning applications it may not correspond to time at all. For example, using a rotating distance sensor (e.g. radar) each time-step corresponds to a different angle of the antenna.

8 A key limitation of a neural network is that the number of inputs and outputs must  
9 be known at training time and must always be the same for all cases. This is not true  
10 for natural languages: if a problem involves processing a sentence, then each input  
11 will be made up of a varying number of words. Similarly for the output, in a text  
12 generation case.

13 Recurrent neural networks (RNN) overcome this by allowing the network to have  
14 a state that persists over time. Inputs of any size can be handled one fixed-sized part  
15 of the input (e.g. one word, or one frame from a video) at a time, using the state to  
16 remember the past inputs.

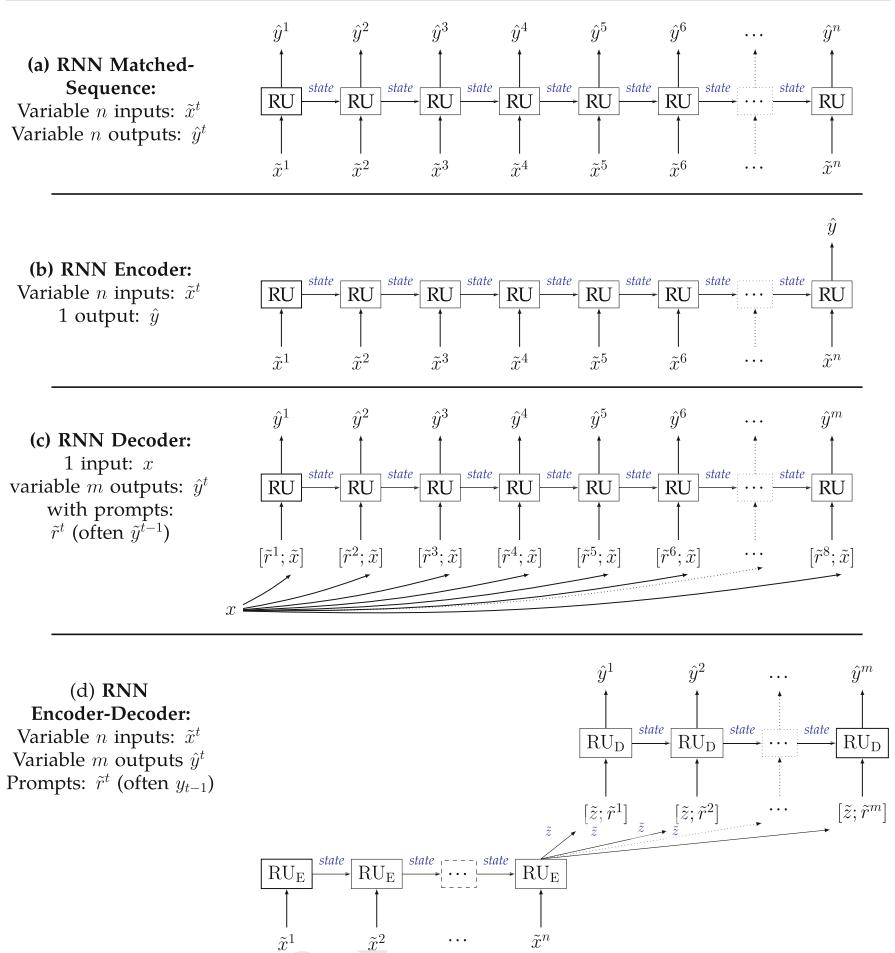
17 A RNN is effectively a chain of feed-forward neural networks, each one being  
18 identical in terms of their weight and bias parameters. While identical in terms of  
19 parameters, they each act at a different time-step.

20 At each time-step the same network is used, with different inputs. For purposes  
21 of looking at this in the big picture, we will first consider each network as a black-  
22 box recurrent unit (RU). The recurrent unit takes at each time-step, an input for that  
23 time-step, some representation of the state for the RU at the previous time-step; and  
24 produces an output for this time-step, and the state representation to be used in the  
25 next time-step. A diagram of this is shown in Fig. 2.2. It is worth distinguishing that  
26 the unit output is not the same as the overall network output, it is just the output of  
27 this sub-network at this time-step. Each time-step can be considered as having two  
28 effective inputs (the previous state, and the actual input) and two effective outputs  
29 (next state and actual output).

30 Not all of these inputs and outputs are actually used at all time-steps meaningfully.  
31 The initial state for the first time-step is normally set to some zero vector, and the  
32 final state at the end of the sequence is normally discarded.

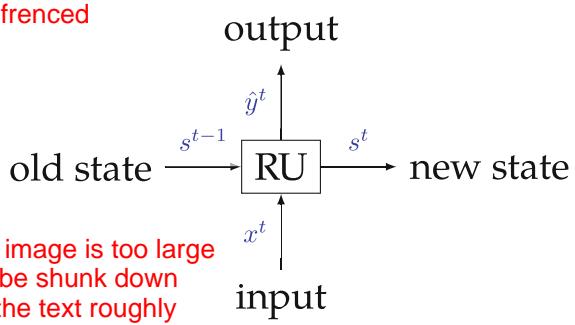
## 33 2.2 General RNN Structures

34 In general most popular uses of recurrent networks belong to one of four general  
35 types of structure. Matched-sequence, encoder, decoder, and encoder-decoder. These  
36 common structures are shown in Fig. 2.1. The motivation for using an RNN comes



**Fig. 2.1** The unrolled structure of an RNN for **a** Matched-sequence **b** Encoding, **c** Decoding and **d** Encoding-Decoding (sequence-to-sequence) problems. RU is the recurrent unit – the neural network which reoccurs at each time-step

**Fig. 2.2** A recurrent unit with its 2 inputs and 2 outputs. Not shown are the internal functioning which may be a complex (e.g. as in LSTM) or simple (e.g. as in a Basic RU) neural network of its own



LW: This image is too large  
it should be shunk down  
to make the text roughly  
consistant in size with the main text

37 from needing to solve problems where the size of the input and/or the size of the  
38 output is not consistent across all cases.

### 39 **2.2.1 Matched-Sequence**

#### **Memory**

The term memory refers to the state. A RNN can be said to remembers something if the information about that is encoded into its state. One should be clear on the difference between learning/training (done by all NNs over the whole the training data) and remembering (done by RNNs within the processing of a single item, across multiple time-steps, during execution).

40 If the input size and the output size is always the same, then one can use a matched-  
41 sequence RNN structure. This is the most basic RNN structure. At each time-step,  
42 there is an input, and a target output. An example of this, in natural language pro-  
43 cessing, is Part of Speech (POS) tagging. Every word is to be labelled as a noun, a  
44 verb, an adjective etc. This does require “memory” as the context words around the  
45 target word being classified influence the correct POS. Remembering the other words  
46 (from other time-steps) is required to use the context to disambiguate cases where  
47 the same word can potentially occupy multiple different parts of speech depend-  
48 ing on the usage. For example `record` is both a noun and a verb. This particular  
49 example is a good use for a Bidirectional RNN (Sect. 2.4.2), as both the previous  
50 and the following words are useful for determining the POS. A key limitation of the  
51 matched-sequence structure is that the input and output size must be the same. Often  
52 though one would like to process an input of that could be any size, but produce just  
53 a fixed size output.

#### **Overlapping Parts of Speech**

It is nearly ubiquitous that verbs and adjectives have a noun form. Words that only occupy a single part of speech are the exception rather than the norm, in English.

### 54 **2.2.2 Encoder**

55 For example if one is trying to learn a mapping from a textual color name to a  
56 probability distribution in color-space (White et al. 2017), then different descrip-  
57 tions have different numbers of words. One input might be `very light green`,  
58 while another just `orange`. At each time-step one input is provided – being one  
59 of the words in the (potentially multi-word) color name. Another example use is in  
60 sentiment analysis, predicting the sentiment being expressed by a text as positive or

negative. All the outputs at all time-steps, except the last, can be ignored. The output of the final time-step can be connected to a further network with a final output layer giving the overall output. The situation, with a variable number of inputs, but a fixed size of output is described as an encoder network.

### 2.2.3 Decoder

#### Pseudo-tokens

Pseudo-tokens such as `<EOS>` (end of string), `<START>` (start of string), and similar are common in RNN tasks. As mentioned, outputting `<EOS>` is required to be known when the decoder is done outputting. `<START>` can be useful to model things that occur near the start of input; and as an initial prompt. In some tasks other non-word tokens might be inserted also, such as in transcription of recorded speech, a `<PAUSE>` token might be included. Pragmatically, as long as the symbols used to represent these never occur amongst the true word tokens, these can be treated just like regular words by the system.

The reverse is a decoder network. This means a fixed sized input being mapped to a variably sized output. If the system is attempting to learn from a point in the color space to the name of that color (Monroe et al. 2016). For example: (144, 238, 144) very light green (the outputs); but (255, 165, 0) might map to just orange (one output). In this decoder type network, there is one true input, at the first time-step, and the output from every time-step is used. Each output can be connected to a softmax layer giving a probability for possible words. Even though the decoder only has a fixed sized input, never-the-less an input must still be provided at every time-step. We call this input a prompt, as it is not providing new information to the network, merely driving it to produce the next output. It is common to include an end of string marker token (often literally `<EOS>`), so as to know when to stop prompting for additional outputs.

#### 2.2.3.1 Prompts in Decoder RNNs

##### Terminology: Prompt

The word prompt is our own terminology. We are not aware of a consistent term used in the literature for the input at each time-step to a decoder RNN. The phrase “dummy input” would also work, although as discussed the choice of prompt can allow useful information to be added. Simplifying the learning problem.

Prompts are required for decoders as the network must have some input to cause it to give an output. In theory this could be a constant: all inputs and all outputs for those inputs, can be learned and remembered by the networks memory (encoded in its state). In practice it is very common to include part of the output of the previous step as part of the prompt. When generating a sequence of words for example, one can use the previous word. At training time this can be the targeted previous output ( $y^{i-1}$ ). At test time (and in real deployment) this is normally the most-likely predicted output ( $\text{argmax } \hat{y}^{t-1}$ ). This effectively gives direct bigram state information to the model, allowing the memory to focus on higher level tasks. It also allows the outputs to be explored, for example by providing the second most-likely word as the first prompt, a different sequence can be generated.

It is also common to include in prompts the original input to the decoder. For example in a caption generator, including a vector representation of the image (for example an Inception Image Embedding Szegedy et al. 2015); in the color decoder example this would be including the original color representation. In general the prompt can be used to add information to the network ensuring that each time-step can do as well as possible, even if the state does not capture all the desired information. Ensuring the state is able to capture all information is part of designing the internals of the RU.

#### 2.2.4 Encoder-Decoder

The encoder-decoder RNN is the generalised structure for sequence to sequence learning (Cho et al. 2014b), it is sometimes called a seq2seq model. It should be contrasted to the matched-sequence RNNs takes a sequence as an input and produces a sequence of *the same size* as an output. The restriction to the sequences being the size limits to the of the matched-sequence RNN for many tasks. For example, in an application such as machine translation, question answering, or video captioning, a sequential input and output is required, but the sequences have different lengths. A sentence in one language will not normally translate to a sentence with the exact same number of word in another. The solution to this is the encoder-decoder RNN. In this structure an encoder RNN is used to take the input, and its final output is then connected as the input to a separate decoder RNN. Thus separating the input processing from the output generation.

#### Output/Input layers

Extra feed-forward layers are often applied between the network's overall input and output, and the RU's input an output. The extra layers could be considered as part of the RU, as they occur at each time-step (even if not used). Alternatively they can be considered as something that surrounds the RU, only at certain time-steps. This depends on the point of view. For our purposes we will consider the core of the RU in isolation.

111 **2.3 Inside the Recurrent Unit LW: Fig 2.2 should be somewhere near here**

112 In this section we discuss the various different types of Recurrent Unit (RU). The  
 113 type of RU determines the different types of recurrent networks, such as Elman  
 114 networks, GRU networks and LSTM networks. As discussed, and shown in Fig. 2.2,  
 115 every recurrent unit from the outside has the previous state, the next state, the unit  
 116 input and the unit output. What differs is how they are connected and what controls  
 117 the information within them. Every recurrent unit is itself a neural network.

118 In many types of RU (e.g. GRU, Basic RU) the output and the state are always  
 119 equal. This particularly makes sense when they should be capturing the same kinds  
 120 of information (as in a decoder-encoder). Furthermore, as there will be additional  
 121 feed-forward layers on top of used outputs (if nothing else an output layer is normally  
 122 required), the need to differentiate output from state is lessened. However, it is a dis-  
 123 tinction made in the very well known LSTM unit (Sect. 2.3.3) so we preserve it here.

124 **2.3.1 Basic Recurrent Unit**

LW: this box should be at the end of this subsection

**Jordan RU**

The Jordan Network formulation is not commonly used today, however analogously to considering the Basic RU as an Elman RU, a similar formulation can be done for a Jordan network (Jordan and MI 1986). The Jordan RU would be given by:

$$\begin{aligned}\tilde{h}^t &= \varphi \left( W [\tilde{x}^t; \tilde{s}^{t-1}] + \tilde{b} \right) \\ \tilde{o}^t &= \varphi \left( V \tilde{h}^t + \tilde{c} \right) \\ \tilde{s}^t &= \tilde{o}^t \\ \hat{y}^t &= \tilde{o}^t\end{aligned}$$

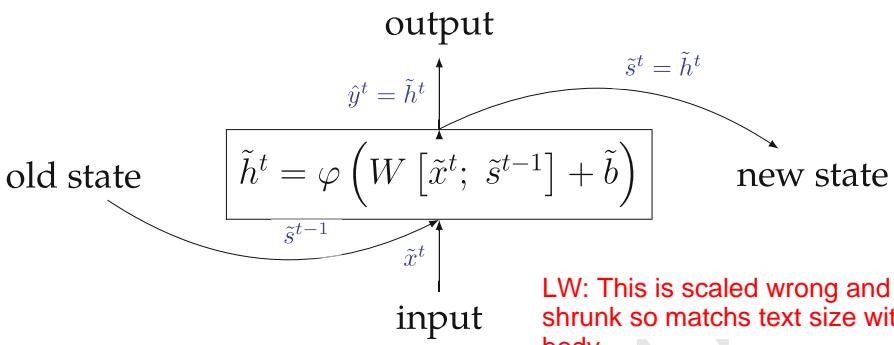
One of these RUs alone is a Jordan network (it does not need an additional output layer). As an RU the difference is in the additional layer  $\tilde{o}^t$  without direct access to the previous state.

125 The most basic recurrent unit, is a single layer  $\tilde{h}^t$ , with the unit output, and the unit  
 126 state both being the value of this layer. The layer is not hidden from the perspective  
 127 of the recurrent unit, but is from the perspective of the whole network. The network  
 128 inside the basic recurrent unit, shown in Fig. 2.3, is given by:

$$\tilde{h}^t = \varphi \left( W [\tilde{x}^t; \tilde{s}^{t-1}] + \tilde{b} \right) \quad (2.1)$$

$$\tilde{s}^t = \tilde{h}^t \quad (2.2)$$

$$\hat{y}^t = \tilde{h}^t \quad (2.3)$$



**Fig. 2.3** A basic recurrent unit.

133 where  $W$  and  $\tilde{b}$  are the weight matrix and bias vector obtainable from training.

134 One could call this an Elman RU: the networks considered in Elman and Jeffrey  
135 (1990) are such a basic RU, with an extra overall output layer on top.

136 Basic RUs are not used in many modern works. It is difficult for these networks  
137 to propagate error information across many time-steps (Bengio et al. 1994). This  
138 results in the state not capturing information from many time-steps ago. Thus the  
139 network has a very short memory.

140 Other more advanced recurrent units solve this problem by placing more explicit  
141 controls on the state.

LW: Jordan RU box should be here

### 142 2.3.2 Gated Recurrent Unit

143 The Gated Recurrent Unit (GRU) was introduced by Cho et al. (2014a). The GRU  
144 is actually a simplification of the much older and better known Long Short Term  
145 Memory (LSTM), which will be discussed in Sect. 2.3.3. We discuss the GRU first  
146 as it is the simpler system. In the evaluations of Chung et al. (2014) and Jozefowicz  
147 et al. (2015) it was found to perform very well – similar to LSTM for most tasks.

The core insight is to gate the changes in the state. There are subnetworks within  
the RU that we call gates. These subnetworks learn how the state should change.

**AQ1** Thus helping the overall network preserve better information in the state (Fig. 2.4)

The gated recurrent unit (GRU) is defined by:

, as shown in Fig. 2.4,

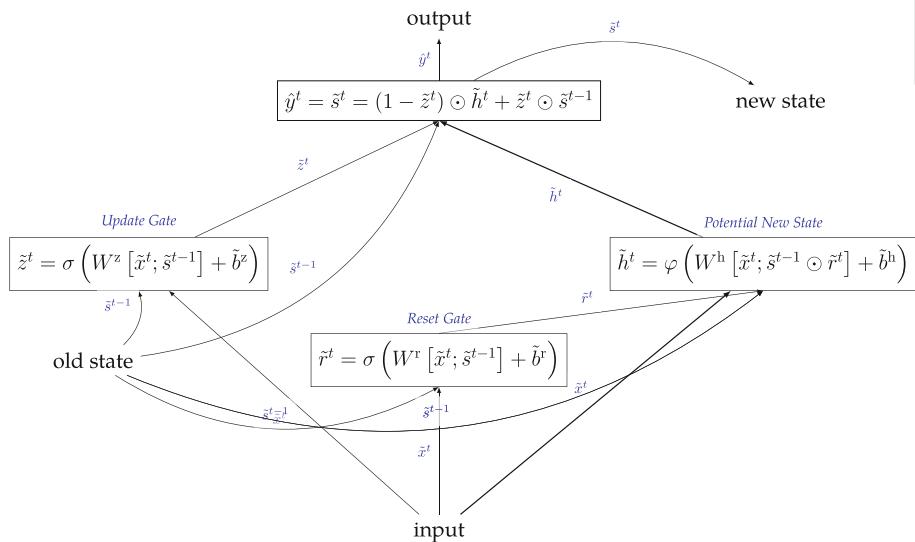
$$152 \quad z^t = \sigma(W^z [\tilde{x}^t; \tilde{s}^{t-1}] + \tilde{b}^z) \quad (2.4)$$

$$153 \quad \tilde{r}^t = \sigma(W^r [\tilde{x}^t; \tilde{s}^{t-1}] + \tilde{b}^r) \quad (2.5)$$

$$154 \quad \tilde{h}^t = \varphi(W^h [\tilde{x}^t; \tilde{s}^{t-1} \odot \tilde{r}^t] + \tilde{b}^h) \quad (2.6)$$

$$155 \quad \tilde{s}^t = (1 - \tilde{z}^t) \odot \tilde{h}^t + \tilde{z}^t \odot \tilde{s}^{t-1} \quad (2.7)$$

$$156 \quad \hat{y}^t = \tilde{s}^t \quad (2.8)$$



**Fig. 2.4** A gated recurrent unit

158 where  $W^z, W^r$  and  $W^h$  are the weight-vectors, and  $\tilde{b}^z, \tilde{b}^r$  and  $\tilde{b}^h$  are the bias vectors  
 159 for the 3 separate layers.

### Trade-off formula

It may be stating the obvious, but a standard way (and the only sensible way) to trade-off a value between two possible choices:  $a$  and  $b$ , when given a preference level for the first as  $p$  (a number between zero and one) is  $(p)a + (1 - p)b$ . If  $p$  is interpreted as the probability of  $a$  being the correct value of a random variable, with  $b$  as the other possible value, then this is the expected value of that random variable.

160 This may look complex, but it can be broken down into parts. First we can see  
 161 that, like in the basic RU, the output and the state have the same value (Eq. 2.8).

162 The  $\tilde{h}^t$  (Eq. 2.6) is our core layer, as in the basic cell. However, unlike in the  
 163 basic cell, it does not immediately become the state:  $\tilde{s}^t$ . There is a trade-off (Eq. 2.7)  
 164 between the state keeping its old value  $\tilde{s}^{t-1}$ , and getting the hidden layer value  $\tilde{h}_t$ .  
 165 This trade-off is element-wise controlled by  $\tilde{z}^t$  which is valued between 0 and 1  
 166 (due to the sigmoid unit). When an element of  $\tilde{z}^t$  is 1, then the value of  $\tilde{s}^{t-1}$  for  
 167 the corresponding element is kept. Conversely, when the  $\tilde{z}^t$  element is 0, then the  
 168 element of the new state is fully given by  $\tilde{h}^t$ .

169 The  $\tilde{z}^t$  is often called the *update gate* as it controls (“gates”) how much the state  
 170 is updated using the new value in  $\tilde{h}^t$ . The update-gate sub-network uses the previous  
 171 state  $\tilde{s}^{t-1}$  and the present input  $\tilde{x}^t$  to make this determination.

### Multiplication, Concatenation, and Addition

It is important to grasp that the product of a matrix and the concentration of two vectors can also be expressed as the sum of the product of a block of that matrix. They are the same thing.  $W \cdot [\tilde{a}; \tilde{b}] = U\tilde{a} + V\tilde{b}$  if  $W = [U \ V]$  The difference is purely notation.

The *reset gate* is loosely similar to the update gate.  $\tilde{r}^t$  controls how much influence the past state  $\tilde{s}^{t-1}$  has on calculating the new value of  $\tilde{h}^t$  – which is the new potential state/output as discussed. It is perhaps clearer if  $\tilde{h}^t$  is reformulated to split  $V$  into the terms which multiply with  $\tilde{x}^t$  and the terms which multiply with  $\tilde{s}^{t-1}$ .

For  $W^h = [W^{hx} \ W^{hs}]$  we can write:

$$\tilde{h}^t = \varphi \left( W^h [\tilde{x}^t; \tilde{s}^{t-1} \odot \tilde{r}^t] + \tilde{b}^h \right) \quad (2.9)$$

$$= \varphi \left( W^{hx} \tilde{x}^t + W^{hs} (\tilde{s}^{t-1} \odot \tilde{r}^t) + \tilde{b}^h \right) \quad (2.10)$$

The  $\tilde{r}^t$  is called the reset-gate because it wipes the effect of the old-state in calculating the potential new state  $\tilde{h}^t$ . When  $\tilde{r}^t$  is reduced to zero, then the updated value for  $\tilde{h}^t$  is as if  $\tilde{s}^{t-1}$  was just like for the initial zero-vector state – it is zeroed out by  $\tilde{r}^t$ . If the update gate  $\tilde{z}^t$  is high then it would fully reset the system.

Notice that when  $\tilde{z}^t = 0$  and  $\tilde{r}^t = 1$ , then the system is identical to the Basic RU. This could be achieved by setting suitably large biases. However, the system is more flexible than that, since  $\tilde{z}^t$  and  $\tilde{r}^t$  are themselves very similar to Basic RUs – though they do not control their own state. The gates can behave differently based on the inputs and states to recognise the most important information that must be stored.

### 2.3.3 LSTM Recurrent Unit

The LSTM is the most well known RNN unit. The term is very nearly interchangeable with RNN today. The original form was proposed by Hochreiter et al. (1997). The form in current use is a variant from Gers et al. (1999).

LSTM uses a compound state, comprised of the units previous output, and an additional state vector called the *cell*. We write  $\tilde{s}^t = (\tilde{c}^t, \tilde{y}^t)$ , to fit the normal formulation, though for maths it is easier to work with these in parts.

#### LW: this should be at the bottom of this section

#### Biasing the Forget Gate

Practically, it is important to initially set the forget gate bias to 1. This is unlike all the other initialisation in the networks to 0 (or a small random number e.g. `randn(0.01)`). Equivalently one can also just add a constant additional bias term of +1 to the forget gate equations. This has been shown to benefit almost all uses of LSTM (Gers et al. 1999; Jozefowicz et al. 2015). This is the default in most neural network libraries (though only recently patched in some).

As shown in Fig. 2.5, the LSTM RU is defined by:

$$\tilde{s}^t = (\tilde{c}^t, \hat{y}^t) \quad (2.11)$$

$$\tilde{i}^t = \sigma(W^i[\tilde{x}^t; \hat{y}_{t-1}] + \tilde{b}^i) \quad (2.12)$$

$$\tilde{f}^t = \sigma(W^f[\tilde{x}^t; \hat{y}_{t-1}] + \tilde{b}^f) \quad (2.13)$$

$$\tilde{o}^t = \sigma(W^o[\tilde{x}^t; \hat{y}_{t-1}] + \tilde{b}^o) \quad (2.14)$$

$$\tilde{h}^t = \tanh(W^h[\tilde{x}^t; \hat{y}_{t-1}] + \tilde{b}^h) \quad (2.15)$$

$$\tilde{c}^t = \tilde{i}^t \odot \tilde{h}^t + \tilde{f}^t \odot \tilde{c}^{t-1} \quad (2.16)$$

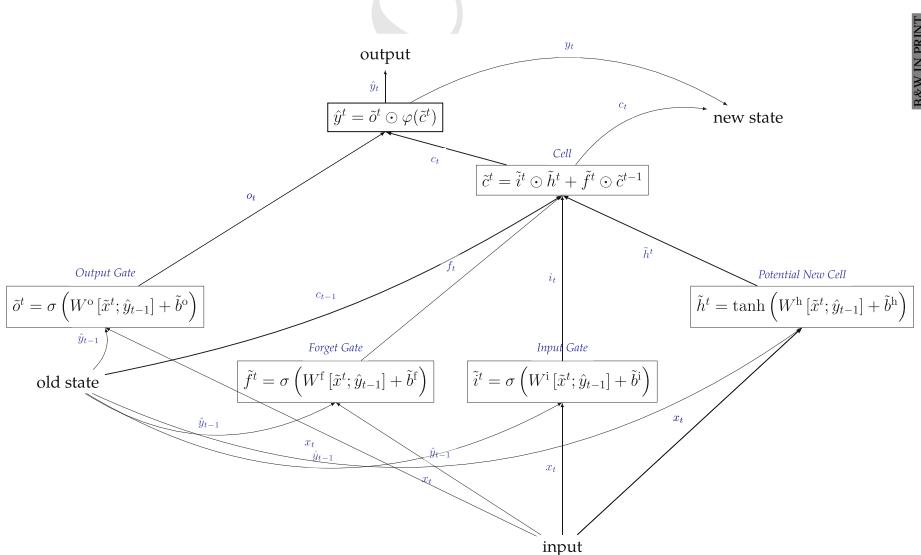
$$\hat{y}^t = \tilde{o}^t \odot \varphi(\tilde{c}^t) \quad (2.17)$$

In LSTM there are three gate sub-networks: the input gate  $\tilde{i}^t$ , the forget gate  $\tilde{f}^t$ , and the output gate  $\tilde{o}^t$ .

Together the input and forget-gates take the purpose of the GRU's update-gate. Consider the case if  $\tilde{o}^t = 1$  and  $\varphi$  is the identity function, and with  $\tilde{i}^t = 1 - \tilde{f}^t$ , that would make the value for  $\tilde{c}^t$  very similar to GRU's  $\tilde{s}^t$  (and  $\hat{y}^t$  identical).

Individually, the forget gate  $\tilde{f}^t$  controls the extent that the previous value of the cell is used, and the input gate  $\tilde{i}^t$  controls the extent to which the new potential value  $\tilde{h}^t$  is used for the new value of  $\tilde{c}^t$ .

The output-gate's obvious purpose is to gate the output  $y_t$ . However, as the output forms part of the state, this has an effect on the networks next time-step. Loosely



B&amp;W IN PRINT

**Fig. 2.5** A LSTM recurrent unit



214 this can be seen as similar to the GRU's reset gate. If we substitute the value for  
 215  $\hat{y}_{t-1} = \tilde{o}^{t-1} \odot \varphi(\tilde{c}^{t-1})$  into  $\tilde{h}^t$  (Fig. 2.5):

$$216 \quad \tilde{h}^t = \tanh \left( W^h [\tilde{x}^t; \hat{y}_{t-1}] + \tilde{b}^h \right) \quad (2.18)$$

$$217 \quad \tilde{h}^t = \tanh \left( W^h [\tilde{x}^t; \tilde{o}^{t-1} \odot \varphi(\tilde{c}^{t-1})] + \tilde{b}^h \right) \quad (2.19)$$

219 it can be seen that, at the previous time-step, setting an element of  $o_{t-1}$  to zero  
 220 effectively removes the effect of the previous cell-state from the equations for all  
 221 the gates and the the potential new cell  $\tilde{h}^t$ . Thus resetting the network. (Similar  
 222 substitutions can be done for  $\tilde{i}^t$ ,  $\tilde{f}^t$  and  $\tilde{o}^t$ )

LW: Bias the Forget gate box should be here!

## 223 2.4 Further Variants

### 224 2.4.1 Deep Variants

225 One can have a deep RNN. This is done by stacking additional recurrent units above  
 226 the existing ones. Attaching, the unit output as the unit input of the layer above. The  
 227 result can be interpreted as a single, deep recurrent unit.

### 228 2.4.2 Bidirectional RNNs

229 A Bidirectional RNN is similar to deep variants (Schuster et al. 1997). This is effec-  
 230 tively using two RNNs so that the input can be processed from both temporal direc-  
 231 tions. The forwards and backwards RNNs both take the same input at each time-step,  
 232 but are otherwise fully distinct. The unit outputs at each time-step are concatenated  
 233 to give an overall time-step output from the RUs.

234 This is not so useful in most decoder RNNs, the number of outputs (and thus time-  
 235 steps) of a decoder RNN is not known except by running them until an end-marker  
 236 is output. However, there are absolutely no issues when using them for an encoder.  
 237 In an encoder network the whole input is available before the output needs to be  
 238 produced. Similarly for a matched-sequence RNN, if the sequence can be broken  
 239 down into blocks (e.g. sentences), and a real-time output is not required.

### 240 2.4.3 Other RNNs

241 There also exist many other RNNs. Some are similar in structure to those discussed  
 242 and can be analysed similarly via their recurrent units. Such as the two models of  
 243 Mikolov et al. (2014), which incorporate a decaying sum of past states into the  
 244 current state (either gated, or constant). Other networks incorporate various other  
 245 data-structures (or analogies to such structures) into their recurrent structure: such

246 as stacks (Dyer et al. 2015), pointers/arrays (Weston et al. 2014), or tapes (Graves  
247 et al. 2014).

248 As natural language is primarily a sequential task, RNNs in various forms will be  
249 discussed in almost every chapter of this book.

250

---

**251 References**

- 252 Bengio, Yoshua, Patrice Simard, and Paolo Frasconi. 1994. Learning long-term dependencies with  
253 gradient descent is difficult. *IEEE Transactions on Neural Networks* 5 (2): 157–166.
- 254 Cho, Kyunghyun, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014a. On the  
255 properties of neural machine translation: Encoder-decoder approaches. In *Eighth workshop on  
256 syntax, semantics and structure in statistical translation (SSST-8)*.
- 257 Cho, Kyunghyun, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger  
258 Schwenk, and Yoshua Bengio. 2014b. Learning phrase representations using RNN encoder-  
259 decoder for statistical machine translation. In *Proceedings of the 2014 conference on empirical  
260 methods in natural language processing (EMNLP)*. Doha, Qatar: Association for Computational  
261 Linguistics, 1724–1734.
- 262 Chung, Junyoung, Caglar Gulcehre, Kyung Hyun Cho, and Yoshua Bengio. 2014. Empirical eval-  
263 uation of gated recurrent neural networks on sequence modeling. [arXiv:1412.3555](https://arxiv.org/abs/1412.3555).
- 264 Dyer, Chris, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A. Smith.  
265 2015. Transition-based dependency parsing with stack long short-term memory. In *CoRR*.  
266 [arXiv:1505.08075](https://arxiv.org/abs/1505.08075).
- 267 Elman, Jeffrey L. 1990. Finding structure in time. *Cognitive Science* 14 (2): 179–211.
- 268 Gers, Felix A., Jürgen Schmidhuber, and Fred Cummins. 1999. Learning to forget: Continual pre-  
269 diction with LSTM. *Neural Computation* 12 (10): 2451–2471
- 270 Graves, Alex, Greg Wayne, and Ivo Danihelka. 2014. Neural turing machines. In *CoRR*.  
271 [arXiv:1410.5401](https://arxiv.org/abs/1410.5401).
- 272 Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*  
273 9 (8): 1735–1780.
- 274 Jordan, M.I. 1986. Serial order: A parallel distributed processing approach (Technical Report No.  
275 8604). San Diego, La Jolla, CA: Institute for Cognitive Science, University of California.
- 276 Jozefowicz, Rafal, Wojciech Zaremba, and Ilya Sutskever. 2015. An empirical exploration of recur-  
277 rent network architectures. In *Proceedings of the 32nd international conference on machine  
278 learning (ICML-15)*, 2342–2350.
- 279 Mikolov, Tomas, Armand Joulin, Sumit Chopra, Michaël Mathieu, and Marc’Aurelio Ranzato.  
280 2014. Learning longer memory in recurrent neural networks. In *CoRR*. [arXiv:1412.7753](https://arxiv.org/abs/1412.7753).
- 281 Monroe, W., N. D. Goodman, and C. Potts. 2016. Learning to generate compositional color descrip-  
282 tions. [arXiv:1606.03821](https://arxiv.org/abs/1606.03821) [cs.CL].
- 283 Schuster, Mike, and Kuldeep K. Palwal. 1997. Bidirectional recurrent neural networks. *IEEE Trans-  
284 actions on Signal Processing* 45 (11): 2673–2681.
- 285 Szegedy, Christian, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov,  
286 Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convo-  
287 lutions. In *The IEEE conference on computer vision and pattern recognition (CVPR)*.
- 288 Weston, Jason, Sumit Chopra, and Antoine Bordes. 2014. Memory networks. In *CoRR*.  
289 [arXiv:1410.3916](https://arxiv.org/abs/1410.3916).
- 290 White, L., R. Togneri, W. Liu, and M. Bennamoun. 2017. Learning distributions of meant color.  
291 [arXiv:1709.09360](https://arxiv.org/abs/1709.09360) [cs.CL].

# Author Queries

Chapter 2

Query Refs.	Details Required	Author's response
AQ1	Please note that the figures are renumbered to ensure sequential order of citations. Please check and confirm the change.	
AQ2	Please check and confirm if the inserted citation of Figs. 2.4 and 2.5 are correct. If not, please suggest an alternate citation.	

LW:

AQ2: Alternates suggested inline

AQ1: Seems correct to me

# Metadata of the chapter that will be visualized in SpringerLink

Book Title	Neural Representations of Natural Language	
Series Title		
Chapter Title	Word Representations	
Copyright Year	2019	
Copyright HolderName	Springer Nature Singapore Pte Ltd.	
Corresponding Author	Family Name Particle Given Name Prefix Suffix Role Division Organization Address Email	<b>White</b> <b>Lyndon</b> <b>Togneri</b> <b>Roberto</b> <b>Liu</b> <b>Wei</b> <b>Bennamoun</b> <b>Mohammed</b>
Author	Family Name Particle Given Name Prefix Suffix Role Division Organization Address Email	School of Electrical, Electronics and Computer Engineering The University of Western Australia Perth, WA, Australia LYNDON.WHITE@RESEARCH.UWA.EDU.AU
Author	Family Name Particle Given Name Prefix Suffix Role Division Organization Address Email	School of Electrical, Electronics and Computer Engineering The University of Western Australia Perth, WA, Australia
Author	Family Name Particle Given Name Prefix Suffix Role Division Organization Address Email	School of Computer Science and Software Engineering The University of Western Australia Perth, WA, Australia
Author	Family Name Particle Given Name Prefix Suffix	

Role  
Division School of Computer Science and Software Engineering  
Organization The University of Western Australia  
Address Perth, WA, Australia  
Email

---

Abstract

Word embeddings are the core innovation that has brought machine learning to the forefront of natural language processing. This chapter discusses how one can create a numerical vector that captures the salient features (e.g. semantic meaning) of a word. Discussion begins with the classic language modelling problem. By solving this, using a neural network-based approach, word-embeddings are created. Techniques such as CBOW and skip-gram models (`word2vec`), and more recent advances in relating this to common linear algebraic reductions on co-locations as discussed. The chapter also includes a detailed discussion of the often confusing hierarchical softmax, and negative sampling techniques. It concludes with a brief look at some other applications and related techniques.

---



# Word Representations

3

*You shall know a word by the company it keeps*

J.R. Firth 1957

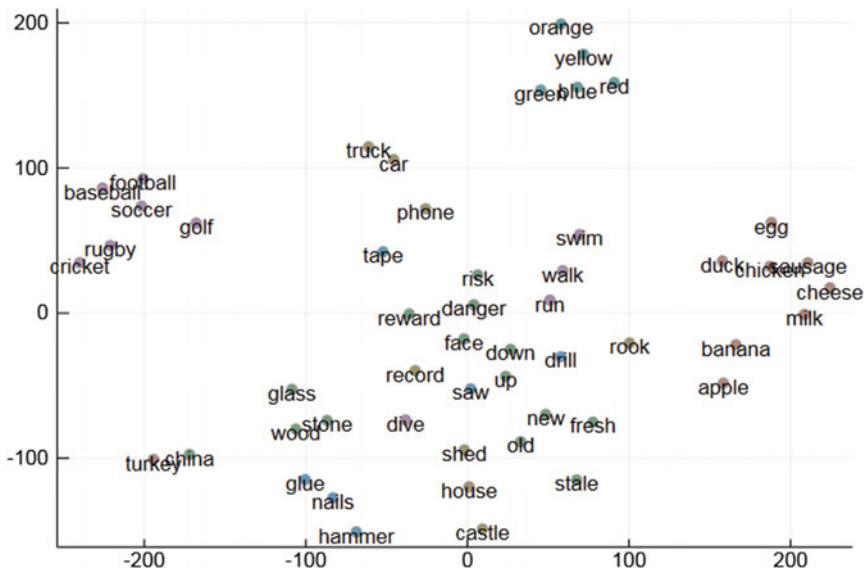
1 Word embeddings are the core innovation that has brought machine learning to the  
2 forefront of natural language processing. This chapter discusses how one can create a  
3 numerical vector that captures the salient features (e.g. semantic meaning) of a word.  
4 Discussion begins with the classic language modelling problem. By solving this,  
5 using a neural network-based approach, word-embeddings are created. Techniques  
6 such as CBOW and skip-gram models (word2vec), and more recent advances in  
7 relating this to common linear algebraic reductions on co-locations as discussed.  
8 The chapter also includes a detailed discussion of the often confusing hierarchical  
9 softmax, and negative sampling techniques. It concludes with a brief look at some  
10 other applications and related techniques.

11 We begin the consideration of the representation of words using neural networks  
12 with the work on language modeling. This is not the only place one could begin the  
13 consideration: the information retrieval models, such as LSI (Dumais et al. 1988)  
14 and LDA (Blei et al. 2003), based on word co-location with documents would be the  
15 other obvious starting point. However, these models are closer to the end point, than  
16 they are to the beginning, both chronologically, and in this chapter's layout. From the  
17 language modeling work, comes the contextual (or acausal) language model works  
18 such as skip-gram, which in turn lead to the post-neural network co-occurrence based  
19 works. These co-occurrence works are more similar to the information retrieval co-  
20 location based methods than the probabilistic language modeling methods for word  
21 embeddings from which we begin this discussion..

AQ1

The epigraph at the beginning of this section is over-used. However, it is  
obligatory to include it in a work such as this, as it so perfectly sums up why





**Fig. 3.1** Some word embeddings from the FastText project (Bojanowski et al. 2016). They were originally 300 dimensions but have been reduced to 2 using t-SNE (Maaten and Geoffrey 2008) algorithm. The colors are from 5 manually annotated categories done before this visualisation was produced: foods, sports, colors, tools, other objects, other. Note that many of these words have multiple meanings (see Chap. 4), and could fit into multiple categories. Also notice that the information captioned by the unsupervised word embeddings is far finer grained than the manual categorisation. Notice, for example, the separation of ball-sports, from words like `run` and `walk`. Not also that `china` and `turkey` are together; this no doubt represents that they are both also countries

representations useful for language modelling are representations that capture semantics (as well as syntax).

### Word Vector or Word Embedding?

Some literature uses the term *word vector*, or *vector-space model* to refer to representations from LDA and LSA etc. Other works use the terms are used synonymously with *word embedding*. Word embeddings are vectors, in any case.

Word embeddings are vector representations of words. An dimensionality reduced scatter plot example of some word embeddings is shown in Fig. 3.1.

## 24 3.1 Representations for Language Modeling

### Probability writing convention

We follow convention that capitalised  $W^i$  is a random variable, and  $w^i$  is a particular value which  $W^i$  may take. The probability of it taking that value would normally be written  $P(W^i=w^i)$ . We simply write  $P(w^i)$  to mean the same thing. This is a common abridged (abuse-of) notation. The random variable in question is implicitly given by the name of its value.

25 The language modeling task is to predict the next word given the prior words (Rosenfeld 2000). For example, if a sentence begins For lunch I will have a  
26 hot, then there is a high probability that the next word will be dog or meal, and  
27 lower probabilities of words such as day or are. Mathematically it is formulated  
28 as:

$$30 P(W^i=w^i \mid W^{i-1}=w^{i-1}, \dots, W^1=w^1) \quad (3.1)$$

31 or to use the compact notation

$$32 P(w^i \mid w^{i-1}, \dots, w^1) \quad (3.2)$$

33 where  $W^i$  is a random variable for the  $i$ th word, and  $w^i$  is a value (a word) it could,  
34 (or does) take. For example:

$$35 P(\text{dog} \mid \text{hot}, \text{a}, \text{want}, \text{I}, \text{lunch}, \text{For})$$

36 The task is to find the probabilities for the various words that  $w^i$  could represent.

### Google n-gram corpora

Google has created a very large scale corpora of 1, 2, 3, 4, and 5-grams from over  $10^{12}$  words from the Google Books project. It has been made freely available at <https://books.google.com/ngrams/datasets> (Lin et al. 2012). Large scale n-gram corpora are also used outside of statistical language modeling by corpus linguists investigating the use of language.

37 The classical approach is trigram statistical language modeling. In this, the number  
38 of occurrences of word triples in a corpus is counted. From this joint probability of  
39 triples, one can condition upon the first two words, to get a conditional probability of  
40 the third. This makes the Markov assumption that the next state depends only on the  
41 current state, and that that state can be described by the previous two words. Under  
42 this assumption Eq. (3.2) becomes:

$$43 P(w^i \mid w^{i-1}, \dots, w^1) = P(w^i \mid w^{i-1}, w^{i-2}) \quad (3.3)$$

44 More generally, one can use an  $n$ -gram language model where for any value of  $n$ ,  
45 this is simply a matter of defining the Markov state to contain different numbers of  
46 words.

47 This Markov assumption is, of-course, an approximation. In the previous example,  
48 a trigram language model finds  $P(w^i | \text{hot}, \text{a})$ . It can be seen that the approximation  
49 has lost key information. Based only on the previous 2 words the next word  $w^i$   
50 could now reasonably be day, but the sentence: For lunch I will have a  
51 hot day makes no sense. However, the Markov assumption in using  $n$ -grams is  
52 required in order to make the problem tractable – otherwise an unbounded amount  
53 of information would need to be stored.

54 A key issue with  $n$ -gram language models is that there exists a data-sparsity  
55 problem which causes issues in training them. Particularly for larger values of  $n$ .  
56 Most combinations of words occur very rarely (Ha et al. 2009). It is thus hard to  
57 estimate their occurrence probability. Combinations of words that do not occur in  
58 the corpus are naturally given a probability of zero. This is unlikely to be true though  
59 – it is simply a matter of rare phrases never occurring in a finite corpus. Several  
60 approaches have been taken to handle this. The simplest is add-one smoothing which  
61 adds an extra “fake” observation to every combination of terms. In common use are  
62 various back-off methods (Katz 1987; Kneser and Hermann 1995) which use the  
63 bigram probabilities to estimate the probabilities of unseen trigrams (and so forth for  
64 other  $n$ -grams.). However, these methods are merely clever statistical tricks – ways  
65 to reassigned probability mass to leave some left-over for unseen cases. Back-off is  
66 smarter than add-one smoothing, as it portions the probability fairly based on the  
67  $(n-1)$ -gram probability. Better still would be a method which can learn to see the  
68 common-role of words (Brown et al. 1992). By looking at the fragment: For lunch  
69 I want a hot, any reader knows that the next word is most likely going to be a  
70 food. We know this for the same reason we know the next word in For elevenses  
71 I had a cold . . . is also going to be a food. Even though elevenses is a  
72 vary rare word, we know from the context that it is a meal (more on this later), and we  
73 know it shares other traits with meals, and similarly have / had, and hot / cold.  
74 These traits influence the words that can occur after them. Hard-clustering words  
75 into groups is nontrivial, particularly given words having multiple meanings, and  
76 subtle differences in use. Thus the motivation is for a language modeling method  
77 which makes use of these shared properties of the words, but considers them in  
78 a flexible soft way. This motivates the need for representations which hold such  
79 linguistic information. Such representations must be discoverable from the corpus,  
80 as it is beyond reasonable to effectively hard-code suitable feature extractors. This  
81 is exactly the kind of task which a neural network achieves implicitly in its internal  
82 representations.

An extended look at classical techniques in statistical language modelling can  
be found in Goodman (2001).

### 83 3.1.1 The Neural Probabilistic Language Model

84 Bengio et al. (2003) present a method that uses a neural network to create a language  
 85 model. In doing so it implicitly learns the crucial traits of words, during training.  
 86 The core mechanism that allowed this was using an embedding or loop-up layer for  
 87 the input.

#### Lookup word embeddings: Hashmap or Array?

The question is purely one of implementation. For purposes of the theory, it does not matter if the implementation is using a String to Vector dictionary (e.g. a hashmap), or a 2D array from which a column is indexed-out (sliced-from) via an integer index representing the word. In the tokenization of the source text, it is common to transform all the words into integers, so as to save memory, especially if string interning is not in use. At that point it makes sense to work with an array. For our notational purposes in this book, we will treat the word  $w^i$  as if it were an integer index, though thinking of it as a string index into a hashmap changes little in the logic.

#### 88 3.1.1.1 Simplified Model Considered with Input Embeddings

89 To understand the neural probabilistic language model, let's first consider a simplified  
 90 neural trigram language model. This model is a simplification of the model introduced  
 91 by Bengio et al. (2003). It follows the same principles, and highlights the most  
 92 important idea in neural language representations. This is that of training a vector  
 93 representation of a word using a lookup table to map a discrete scalar word to a  
 94 continuous-space vector which becomes the first layer of the network.

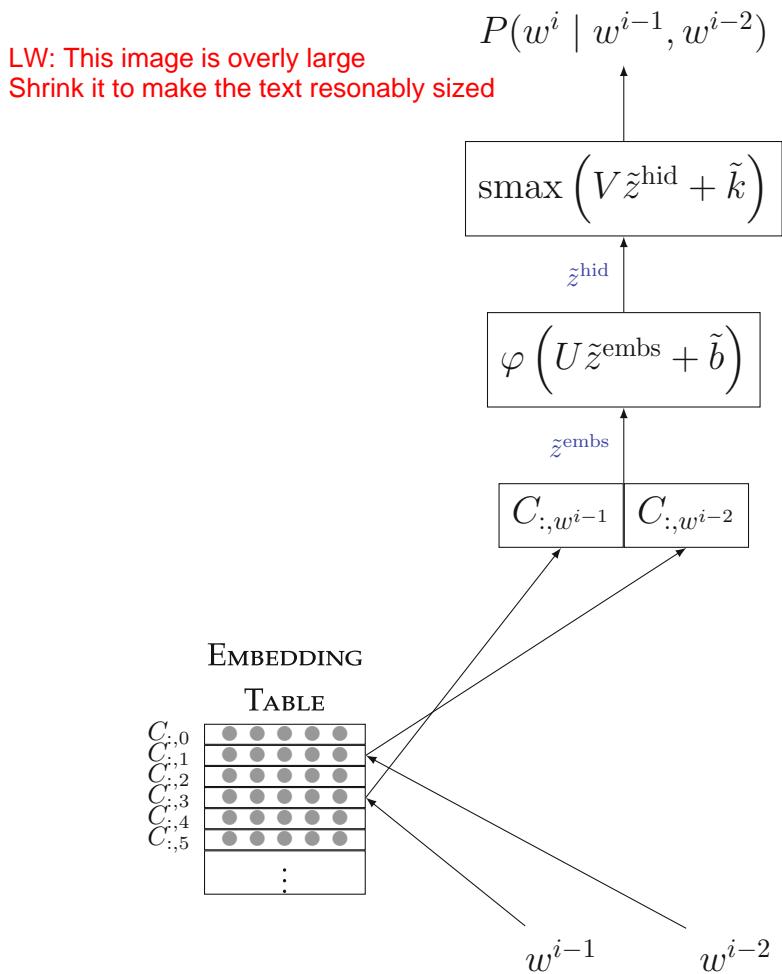
95 The neural trigram probabilistic network is defined by:

$$96 P(w^i | w^{i-1}, w^{i-2}) = \text{smax} \left( V \varphi \left( U [C_{:, w^{i-1}}; C_{:, w^{i-2}}] + \tilde{b} \right) + \tilde{k} \right) \quad (3.4)$$

99 where  $U$ ,  $V$ ,  $\tilde{b}$ ,  $\tilde{k}$  are the weight matrices and biases of the network. The matrix  $C$   
 100 defines the embedding table, from which the word embeddings,  $C_{:, w^{i-1}}$  and  $C_{:, w^{i-2}}$ ,  
 101 representing the previous two words ( $w^{i-1}$  and  $w^{i-2}$ ) are retrieved. The network is  
 102 shown in Fig. 3.2.

$C_{:, w^i}$  not  $C_{:, i}$

Note that here we use the word  $w^i$  as the index to lookup the word embeddings.  $i$  is the index of the word index in the corpus. That is to say that if the  $i$ th word, and the  $j$ th word are the same: i.e  $w^i = w^j$ , then they will index out the same vector from  $C$ .  $w^i = w^j \implies C_{:, w^i} = C_{:, w^j}$ .



**Fig. 3.2** The Neural Trigram Language Model

### One-hot product or Indexed-lookup

In some works you may see the process of retrieving the word vector from a matrix of word vectors described as a one-hot multiplication. For a word represented by the index  $w$ , where  $\tilde{e}^w$  the one-hot vector with a 1 in the  $w$ th position, and for  $C$ , the table of word embeddings, one can write  $C \tilde{e}^w$  to find the embedding for  $w$ . We will write  $C_{:,w}$  and refer to this as looking up the word vector from the  $w$ th column. Of-course  $C_{:,w} = C \tilde{e}^w$ , however in practical implementation the performance ramifications are huge. Matrix column indexing is effectively an  $O(1)$  operation (in a column major languages), whereas a dense matrix-vector product is  $O(n^2)$ . The one-hot product can be

LW: Move this box so it is not split over two pages

used in a pinch to support using embeddings in neural network toolkits that do not support lookup/embedding layers. However, we strongly suggest that if your toolkit does not support lookup/embedding layers then it is unsuitable for use in NLP applications. Some tool-kits, e.g. Flux.jl (<https://github.com/FluxML/Flux.jl>), explicitly handle sparse one-hot types, and automatically make this transformation. In that case, it is outright equivalent.

103 In the neural trigram language model, each of the previous two words is used to  
 104 look-up a vector from the embedding matrix. These are then concatenated to give  
 105 a dense, continuous-space input to the above hidden layer. The output layer is a  
 106 softmax layer, it gives the probabilities for each word in the vocabulary, such that  
 107  $\hat{y}_{w^i} = P(w^i | w^{i-1}, w^{i-2})$ . Thus producing a useful language model.

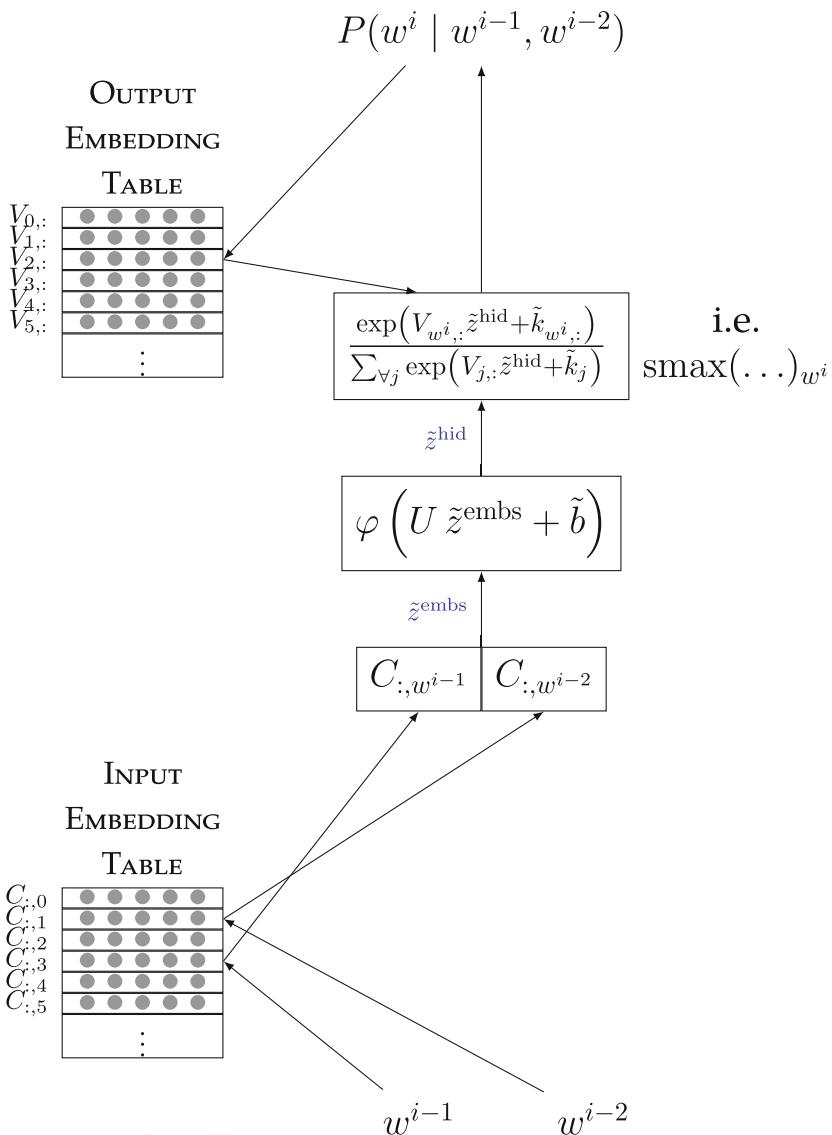
108 The word embeddings are trained, just like any other parameter of the network  
 109 (i.e. the other weights and biases) via gradient descent. An effect of this is that the  
 110 embeddings of words which predict the same future word will be adjusted to be nearer  
 111 to each other in the vector space. The hidden layer learns to associate information  
 112 with regions of the embedding space, as the whole network (and every layer) is a  
 113 continuous function. This effectively allows for information sharing between words.  
 114 If two word's vectors are close together because they mostly predict the same future  
 115 words, then that area of the embedding space is associated with predicting those  
 116 words. If words  $a$  and  $b$  often occur as the word prior to some similar set of words  
 117 ( $w, x, y, \dots$ ) in the training set and word  $b$  also often occurs in the training set before  
 118 word  $z$ , but (by chance)  $a$  never does, then this neural language model will predict  
 119 that  $z$  is likely to occur after  $a$ . Where-as an n-gram language model would not. This  
 120 is because  $a$  and  $b$  have similar embeddings, due to predicting a similar set of words.  
 121 The model has learnt common features about these words implicitly from how they  
 122 are used, and can use those to make better predictions. These features are stored in  
 123 the embeddings which are looked up during the input.

### 124 3.1.1.2 Simplified Model Considered with Input and Output 125 Embeddings

126 We can actually reinterpret the softmax output layer as also having embeddings. An  
 127 alternative but equivalent diagram is shown in Fig. 3.3.

**Consider, that the matrix product of a row vector with a column vector is the dot product**

$V_{w_i, \cdot} \tilde{z}^{\text{hid}}$  can be seen as computing the dot product between the output embedding for  $w_i$  and the hidden layer representation of the prior words/context ( $w^{i-1}$  and  $w^{i-2}$  in this case) in the form of  $\tilde{z}^{\text{hid}}$ . This leads to an alternate interpretation of the whole process as minimising the dot-product distance between the output embedding and the context representation. This is particularly relevant for the skip-gram model discussed in Sect. 3.2.2 (with just one input word and no hidden layer).



**Fig. 3.3** Neural Trigram Language Model as considered with output embeddings. This is mathematically identical to Fig. 3.2.

LW: This image is overly large,  
shink it such that the text is reasonably scaled

128 The final layer of the neural trigram language model can be rewritten per each index  
 129 corresponding to a possible next word ( $w^i$ ):

$$130 \quad \text{smax}(V_{\tilde{z}}^{\text{hid}} + \tilde{k})_{w^i} = \frac{\exp(V_{w^i,:} \tilde{z}^{\text{hid}} + \tilde{k}_{w^i})}{\sum_{\forall j} \exp(V_{j,:} \tilde{z}^{\text{hid}} + \tilde{k}_j)} \quad (3.5)$$

131 In this formulation, we have  $V_{w_i,:}$  as the output embedding for  $w^i$ . As we considered  
 132  $C_{:,w_i}$  as its input embedding.

### 133 3.1.1.3 Bayes-Like Reformulation

134 When we consider the model with output embeddings, it is natural to also consider  
 135 it under the light of the Bayes-like reformulation from Sect. 1.5.1.1:

$$136 \quad P(Y=i \mid Z=\tilde{z}) = \frac{R(Z=\tilde{z} \mid Y=i) R(Y=i)}{\sum_{\forall j} R(Z=\tilde{z} \mid Y=j) R(Y=j)} \quad (3.6)$$

138 which in this case is:

$$139 \quad P(w^i \mid w^{i-1}, w^{i-2}) = \\ 140 \quad \frac{R(Z=\tilde{z}^{\text{hid}} \mid W^i=w^i) R(W^i=w^i)}{\sum_{\forall v \in \mathbb{V}} R(Z=\tilde{z}^{\text{hid}} \mid W^i=v) R(W^i=v)} \quad (3.7)$$

142 where  $\sum_{\forall v \in \mathbb{V}}$  is summing over every possible word  $v$  from the vocabulary  $\mathbb{V}$ , which  
 143 does include the case  $v = w^i$ .

144 Notice the term:

$$145 \quad \frac{R(W^i=w^i)}{\sum_{\forall v \in \mathbb{V}} R(W^i=v)} = \frac{\exp(\tilde{k}_{w^i})}{\sum_{\forall v \in \mathbb{V}} \exp(\tilde{k}_v)} \quad (3.8)$$

$$146 \quad = \frac{1}{\sum_{\forall v \in \mathbb{V}} \exp(\tilde{k}_v - \tilde{k}_{w^i})} \quad (3.9)$$

148 The term  $R(W^i=w^i) = \exp(\tilde{k}_{w^i})$  is linked to the unigram word probabilities:  
 149  $P(Y=y)$ . If  $\mathbb{E}(R(Z \mid W_i)) = 1$  then the optimal value for  $\tilde{k}$  would be given by  
 150 the log unigram probabilities:  $k_{w^i} = \log P(W^i=w^i)$ . This condition is equivalent  
 151 to if  $\mathbb{E}(V\tilde{z}^{\text{hid}}) = 0$ . Given that  $V$  is normally<sup>1</sup> initialized as a zero mean Gaussian,  
 152 this condition is at least initially true. This suggests, interestingly, that we can pre-  
 153 determine good initial values for the output bias  $\tilde{k}$  using the log of the unigram  
 154 probabilities. In practice this is not required, as it is learnt rapidly by the network  
 155 during training.

---

<sup>1</sup>No pun intended.

### 3.1.1.4 The Neural Probabilistic Language Model

Bengio et al. (2003) derived a more advanced version of the neural language model discussed above. Rather than being a trigram language model, it is an  $n$ -gram language model, where  $n$  is a hyper-parameter of the model. The knowledge sharing allows the data-sparsity issues to be ameliorated, thus allowing for a larger  $n$  than in traditional n-gram language models. Bengio et al. (2003) investigated values for 2, 4 and 5 prior words (i.e. a trigram, 5-gram and 6-gram model). The network used in their work was marginally more complex than the trigram neural language model. As shown in Fig. 3.4, it features a layer-bypass connection. For  $n$  prior words, the model is described by:

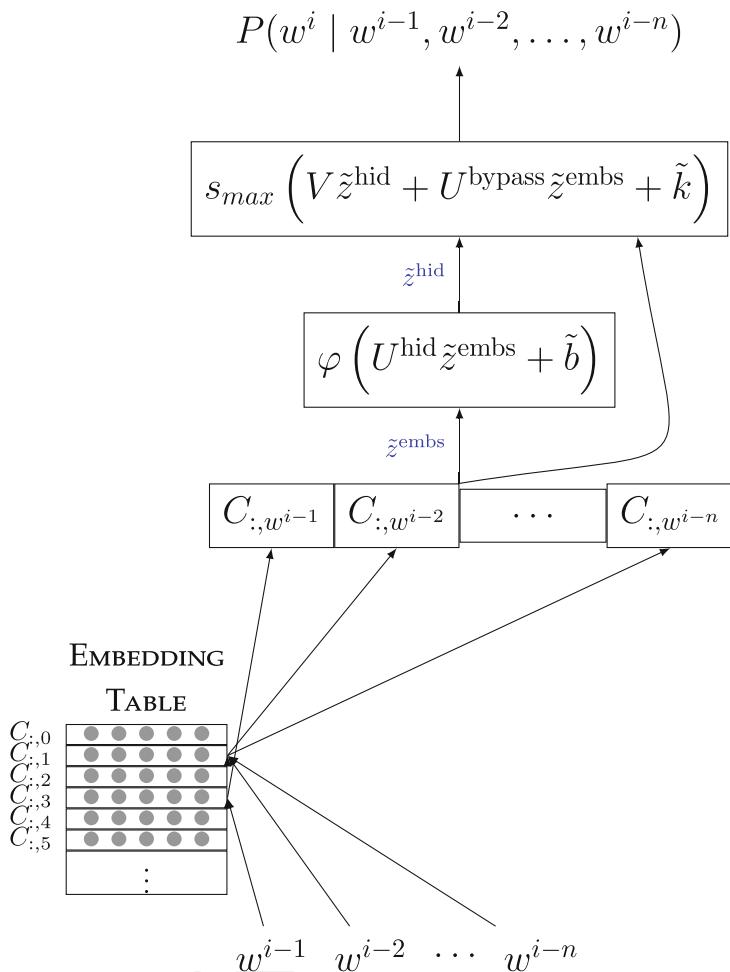
$$\begin{aligned}
 P(w^i | w^{i-1}, \dots, w^{i-n}) = & \text{smax}(\text{LW: realign this to here} \\
 & + V \varphi(U^{\text{hid}} [C_{:, w^{i-1}}; \dots; C_{:, w^{i-n}}] + \tilde{b}) \\
 & + U^{\text{bypass}} [C_{:, w^{i-1}}; \dots; C_{:, w^{i-n}}] \\
 & + \tilde{k})_{w^i}
 \end{aligned} \tag{3.10}$$

The layer-bypass is a connivance to aid in the learning. It allows the input to directly affect the output without being mediated by the shared hidden layer. This layer-bypass is an unusual feature, not present in future works deriving from this, such as Schwenk (2004). Though in general it is not an unheard of technique in neural network machine learning.

#### Input vocabulary and output vocabulary do not have to be the same

Schwenk (2004) suggests using only a subset of the vocabulary as options for the output, while allowing the full vocabulary in the input space – with a fall-back to classical language models for the missed words. This decreases the size of the softmax output layer, which substantially decreases the time taken to train or evaluate the network. As a speed-up technique this is now eclipsed by hierarchical softmax and negative sampling discussed in Sect. 3.4. The notion of a different input and output vocabulary though remains important for word-sense embeddings as will be discussed in Chap. 4.

This is the network which begins the notions of using neural networks with vector representations of words. Bengio et al. focused on the use of the of sliding window of previous words – much like the traditional n-grams. At each time-step the window is advanced forward and the next is word predicted based on the shifted context of prior words. This is of-course exactly identical to extracting all n-grams from the corpus and using those as the training data. They very briefly mention that an RNN could be used in its place.

**Fig. 3.4** Neural Probabilistic Language Model

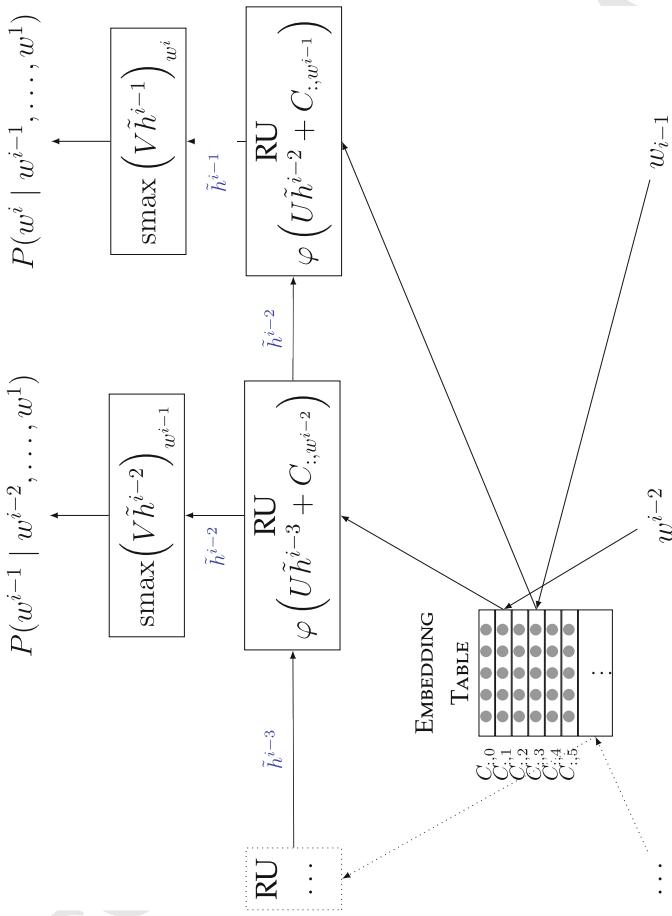
LW: This image is too large, shrink to match text size

### 183 3.1.2 RNN Language Models

184 In Mikolov et al. (2010) an RNN is used for language modelling, as shown in Fig. 3.5.  
 185 Using the terminology of Chap. 2, this is an encoder RNN, made using Basic Recur-  
 186 rent Units. Using an RNN eliminates the Markov assumption of a finite window of  
 187 prior words forming the state. Instead, the state is learned, and stored in the state  
 188 component of the RUs.

**No Bias?** LW: Move box such that it is not split across 2 pages

It should be noticed that Eqs.(3.11) and (3.12) are missing the bias terms.  
 This is not commented on in Mikolov et al. (2010). But in the corresponding



**Fig. 3.5** RNN Language Model. The RU equation shown is the basic RU used in Mikolov et al. (2010). It can be substituted for a LSTM RU or an GRU as was done in Sundermeyer et al. (2012); Jozefowicz et al. (2015), with appropriate changes.

LW Image is too large.  
Resize such that it fits the right way up.

chapter of Mikolov’s thesis (Tomas 2012), it is explicitly noted that biases were not used in the network as it was not found that they gave a significant improvement to the result. This is perhaps surprising, particularly in the output softmax layer given the very unbalanced class (unigram) frequencies.

In the papers for several of Mikolov’s other works, including those for skip-gram and CBOW discussed in Sect. 3.2, the bias terms are also excluded. We have matched those equations here. We do note though, that it is likely that many publicly available implementations of these algorithms would include the bias term: due either to a less close reading of the papers, or to the assumption that the equations are given in *design matrix* form: where the bias is not treated as a separate term to the weights, and the input is padded with an extra 1. We do not think this is at all problematic.

We discuss this further for the case of hierarchical softmax in Sect. 3.4.1, where the level is a proxy for the unigram frequency – and thus for the bias.

189     This state  $\tilde{h}_i$  being the hidden state (and output as this is a basic RU) from the  
 190     *i* time-step. The *i*th time-step takes as its input the *i*th word. As usual this hidden  
 191     layer was an input to the hidden-layer at the next time-step, as well as to the output  
 192     softmax.

$$193 \quad \tilde{h}^i = \varphi \left( U\tilde{h}^{i-1} + C_{:, w_{i-1}} \right) \quad (3.11)$$

$$194 \quad P(w^i | w^{i-1}, \dots, w^1) = \text{smax} \left( V\tilde{h}^{i-1} \right)_{w^i} \quad (3.12)$$

196     Rather than using a basic RU, a more advanced RNN such as a LSTM or GRU-based  
 197     network can be used. This was done by Sundermeyer et al. (2012) and Jozefowicz  
 198     et al. (2015), both of whom found that the more advanced networks gave significantly  
 199     better results.

## 200 3.2 Acausal Language Modeling

201     The step beyond a normal language model, which uses the prior words to predict  
 202     the next word, is what we will term acausal language modelling. Here we use the  
 203     word acausal in the signal processing sense. It is also sometimes called contextual  
 204     language modelling, as the whole context is used, not just the prior context. The task  
 205     here is to predict a missing word, using the words that precede it, as well as the words  
 206     that come after it.

207     As it is acausal it cannot be implemented in a real-time system, and for many  
 208     tasks this renders it less, directly, useful than a normal language model. However, it  
 209     is very useful as a task to learn a good representation for words.

### Are CBOW & Skip-Gram Neural Networks?

It is sometimes asserted that these models are not in-fact neural networks at all. This assertion is often based on their lack of a traditional hidden-layer, and similarities in form to several other mathematical models (discussed in Sect. 3.3). This distinction is purely academic though. Any toolkit that can handle the prior discussed neural network models can be used to implement CBOW and Skip-Gram, more simply than using a non-neural network focused optimiser.

It also should be noted that embedding lookup is functionally an unusual hidden layer – this becomes obvious when considering the lookup as an one-hot product. Though it does lack a non-linear activation function.

210 The several of the works discussed in this section also feature hierarchical softmax  
 211 and negative sampling methods as alternative output methods. As these are compli-  
 212 cated and easily misunderstood topics they are discussed in a more tutorial fashion  
 213 in Sect. 3.4. This section will focus just on the language model logic; and assume the  
 214 output is a normal softmax layer.

#### 3.2.1 Continuous Bag of Words

216 The continuous bag of words (CBOW) method was introduced by Mikolov et al.  
 217 (2013a). In truth, this is not particularly similar to bag of words at all. No more so  
 218 than any other word representation that does not have regard for order of the context  
 219 words (e.g. skip-gram, and GloVe).

220 The CBOW model takes as its input a context window surrounding a central  
 221 skipped word, and tries to predict the word that it skipped over. It is very similar to  
 222 earlier discussed neural language models, except that the window is on both sides.  
 223 It also does not have any non-linearities; and the only hidden layer is the embedding  
 224 layer.

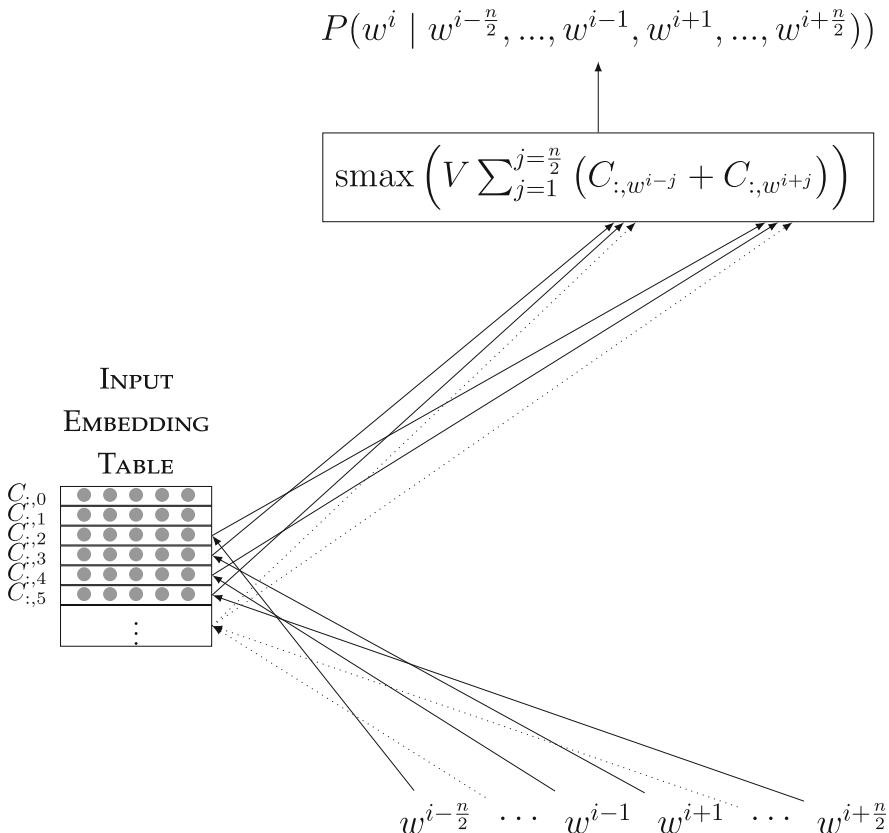
225 For a context window of width  $n$  words – i.e.  $\frac{n}{2}$  words to either side, of the target  
 226 word  $w^i$ , the CBOW model is defined by:

$$\begin{aligned} 227 \quad P(w^i | w^{i-\frac{n}{2}}, \dots, w^{i-1}, w^{i+1}, \dots, w^{i+\frac{n}{2}}) \\ 228 \quad = \text{smax} \left( V \sum_{j=i+1}^{j=\frac{n}{2}} (C_{:,w^{i-j}} + C_{:,w^{i+j}}) \right)_{w^i} \end{aligned} \quad (3.13)$$

230 This is shown in diagrammatic form in Fig. 3.6. By optimising across a training  
 231 dataset, useful word embeddings are found, just like in the normal language model  
 232 approaches (Fig. 3.7).

AQ2





**Fig. 3.6** CBOW Language Model

### 3.2.2 Skip-Gram

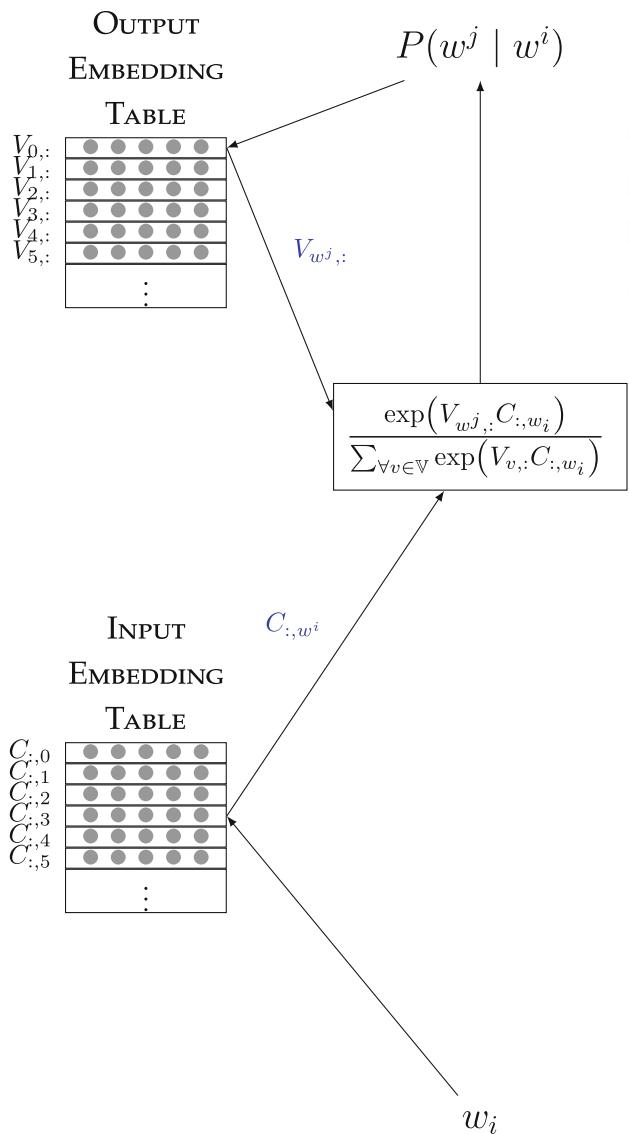
#### Skip-gram naming

In different publications this model may be called skipgram, skip-gram, skip-ngram, skip gram etc. Further, it may be called word2vec after the publicly released implementation of the algorithm. Though the word2vec software can also be used for CBOW, so sometimes it can refer to CBOW.

234 The converse of CBOW is the skip-grams model (Mikolov et al. 2013a). In this  
 235 model, the central word is used to predict the words in the context.

236 The model itself is single word input, and its output is a softmax for the probability  
 237 of each word in the vocabulary occurring in the context of the input word. This can  
 238 be indexed to get the individual probability of a given word occurring as usual for a  
 239 language model. So for input word  $w^i$  the probability of  $w^j$  occurring in its context  
 240 is given by:





**Fig. 3.7** Skip-gram language Language Model. Note that the probability  $P(w^j \mid w^i)$  is optimised during training for every  $w^j$  in a window around the central word  $w^i$ . Note that the final layer in this diagram is just a softmax layer, written in in output embedding form

241                     $P(w^j | w^i) = \text{smax}(V C_{:,w^i})_{w^j}$                     (3.14)

This is shown in Figure 3.7.

242 The goal, is to maximise the probabilities of all the observed outputs that actually do  
 243 occur in its context. This is done, as in CBOW by defining a window for the context  
 244 of a word in the training corpus,  $(i - \frac{n}{2}, \dots, i - 1, i + i, \dots, i + \frac{n}{2})$ . It should be  
 245 understood that while this is presented similarly to a classification task, there is no  
 246 expectation that the model will actually predict the correct result, given that even  
 247 during training there are multiple correct results. It is a regression to an accurate  
 248 estimate of the probabilities of co-occurrence (this is true for probabilistic language  
 249 models more generally, but is particularly obvious in the skip-gram case).

250 Note that in skip-gram, like CBOW, the only hidden layer is the embedding layer.  
 251 Rewriting Eq. (3.14) in output embedding form:

252                     $P(w^j | w^i) = \text{smax}(V C_{:,w^i})_{w^j}$                     (3.15)

253                     $P(w^j | w^i) = \frac{\exp(V_{w^j,:} C_{:,w^i})}{\sum_{v \in \mathbb{V}} \exp(V_{v,:} C_{:,v})}$                     (3.16)

255 The key term here is the product  $V_{w^j,:} C_{:,w^i}$ . The remainder of Eq. (3.16) is to normalise this into a probability. Maximising the probability  $P(w^j | w^i)$  is equivalent  
 256 to maximising the dot produce between  $V_{w^j,:}$ , the output embedding for  $w^j$  and  
 257  $C_{:,w^i}$  the input embedding for  $w^i$ . This is to say that the skip-gram probability is  
 258 maximised when the angular difference between the input embedding for a word,  
 259 and the output embeddings for its co-occurring words is minimised. The dot-product  
 260 is a measure of vector similarity – closely related to the cosine similarity.  
 261

Skip-gram is much more commonly used than CBOW.

### 263 3.2.3 Analogy Tasks

#### Analogy Tasks uncover prejudice in corpora LW: Move to end of section

Bolukbasi et al. (2016) and Caliskan et al. (2017) use analogy tasks, and related variant formulations to find troubling associations between words, such as Bolukbasi et. al's titular Man is to Computer Programmer, as Woman is to Homemaker. Finding these relationships in the embedding space, indicated that they are present in the training corpus, which in turn shows their prevalence in society at large. It has been observed that machine learning can be a very good mirror upon society.

264 One of the most notable features of word embeddings is their ability to be used  
 265 to express analogies using linear algebra. These tasks are keyed around answering  
 266 the question:  $b$  is to  $a$ , as what is to  $c$ ? For example, a semantic analogy would be  
 267 answering that Aunt is to Uncle as King is to Queen. A syntactic analogy would be  
 268 answering that King is to Kings as Queen is to Queens. The latest and largest  
 269 analogy test set is presented by Gladkova et al. (2016), which evaluates embeddings

on 40 subcategories of knowledge. Analogy completion is not a practical task, but rather serves to illustrate the kinds of information being captured, and the way in which it is represented (in this case linearly).

The analogies work by relating similarities of differences between the word vectors. When evaluating word similarity using word embeddings a number of measures can be employed. By far the cosine similarity is the most common. This is given by

$$\text{sim}(\tilde{u}, \tilde{v}) = \frac{\tilde{u} \cdot \tilde{v}}{\|\tilde{u}\| \|\tilde{v}\|} \quad (3.17)$$

This value becomes higher the closer the word embedding  $\tilde{u}$  and  $\tilde{v}$  are to each other, ignoring vector magnitude. For word embeddings that are working well, then words with closer embeddings should have correspondingly greater similarity. This similarity could be syntactic, semantic or other. The analogy tasks can help identify what kinds of similarities the embeddings are capturing.

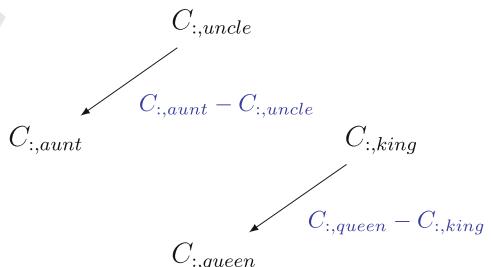
Using the similarity scores, a ranking of words to complete the analogy is found. To find the correct word for  $d$  in:  $d$  is to  $c$  as  $b$  is to  $a$  the following is computed using the table of embeddings  $C$  over the vocabulary  $\mathbb{V}$ :

$$\underset{\forall d \in \mathbb{V}}{\text{argmax}} \text{ sim}(C_{:,d} - C_{:,c}, C_{:,a} - C_{:,b}) \quad (3.18)$$

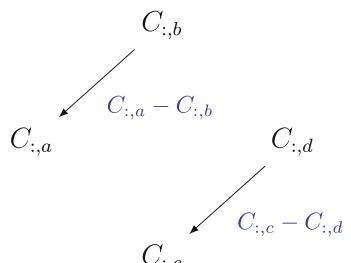
$$\text{i.e. } \underset{\forall d \in \mathbb{V}}{\text{argmax}} \text{ sim}(C_{:,d}, C_{:,a} - C_{:,b} + C_{:,c}) \quad (3.19)$$

This is shown diagrammatically in Figs. 3.8 and 3.9. Sets of embeddings where the vector displacement between analogy terms are more consistent score better.

**Fig. 3.8** Example of analogy algebra



**Fig. 3.9** Vectors involved in analogy ranking tasks, this may help to understand the math in Eq. (3.19)



Initial results in Mikolov et al. (2013c) were relatively poor, but the surprising finding was that this worked at all. Mikolov et al. (2013a) found that CBOW performed poorly for semantic tasks, but comparatively well for syntactic tasks; skip-gram performed comparatively well for both, though not quite as good in the syntactic tasks as CBOW. Subsequent results found in Pennington et al. (2014) were significantly better again for both.

LW: Move Analogy uncover bias box here

### 3.3 Co-location Factorisation

#### 3.3.1 GloVe

##### Distance weighted co-occurrence and dynamic window sizing

When training skip-gram and CBOW, Mikolov et al. used dynamic window sizing. This meant that if the specified window size was  $n$ , in any given training case being considered the actual window size was determined as a random number between 0 and  $n$ .

Pennington et al. achieve a similar effect by weighting co-occurrences within a window with inverse proportion to the distance between the word. That is to say if  $w^i$  and  $w^j$  occur in the same window (i.e.  $|i - j| < n$ ), then rather than contributing 1 to the entry in the co-occurrence count  $X_{w^i, w^j}$ , they contribute  $\frac{1}{|i - j|}$ .

LW: Move the text of these boxes into the bottom box.

##### Subsampling, and weight saturation

Skip-gram and CBOW models use a method called subsampling to decrease the effect of common words. The subsampling method is to randomly discard words from training windows based on their unigram frequency. This is closely related to the saturation of the co-occurrence weights as calculated by  $f(X)$  used by GloVe. Averaged over all training cases the effect is nearly the same.

##### Key Factors Mentioned as Asides LW move to end of section.

There is an interesting pattern of factors being considered as not part of the core algorithm. We have continued this in the side-notes of this section; with the preceding notes on Distance weighting and subsampling. While the original papers consider these as unimportant to the main thrust of the algorithms (Levy et al. 2015) found them to be crucial hyper-parameters.

Skip-gram, like all probabilistic language models, is a intrinsically prediction-based method. It is effectively optimising a neural network to predict which words will co-occur in the with in the range of given by the context window width. That optimisation is carried out per-context window, that is to say the network is updated

303 based on the local co-occurrences. In Pennington et al. (2014) the authors show that  
 304 if one were to change that optimisation to be global over all co-occurrences, then  
 305 the optimisation criteria becomes minimising the cross-entropy between the true  
 306 co-occurrence probabilities, and the value of the embedding product, with the cross  
 307 entropy measure being weighted by the frequency of the occurrence of the word. That  
 308 is to say if skip-gram were optimised globally it would be equivalent to minimising:

$$309 \quad Loss = - \sum_{\forall w^i \in \mathbb{V}} \sum_{\forall w^j \in \mathbb{V}} X_{w^i, w^j} P(w^j | w^i) \log(V_{w^j, :} C_{:, w^i}) \quad (3.20)$$

310 for  $\mathbb{V}$  being the vocabulary and for  $X$  being the a matrix of the true co-occurrence  
 311 counts, (such that  $X_{w^i, w^j}$  is the number of times words  $w^i$  and  $w^j$  co-occur), and  
 312 for  $P$  being the predicted probability output by the skip-gram.

313 Minimising this cross-entropy efficiently means factorising the true co-occurrence  
 314 matrix  $X$ , into the input and output embedding matrices  $C$  and  $V$ , under a particular  
 315 set of weightings given by the cross entropy measure.

316 Pennington et al. (2014) propose an approach based on this idea. For each word  
 317 co-occurrence of  $w^i$  and  $w^j$  in the vocabulary: they attempt to find optimal values  
 318 for the embedding tables  $C$ ,  $V$  and the per word biases  $\tilde{b}$ ,  $\tilde{k}$  such that the function  
 319  $s(w^i, w^j)$  (below) expresses an approximate log-likelihood of  $w^i$  and  $w^j$ .

$$320 \quad \text{optimise } s(w^i, w^j) = V_{w^j, :} C_{:, w^i} + \tilde{b}_{w^i} + \tilde{k}_{w^j} \quad (3.21)$$

$$321 \quad \text{such that } s(w^i, w^j) \approx \log(X_{w^i, w^j}) \quad (3.22)$$

323 This is done via the minimisation of

$$324 \quad Loss = - \sum_{\forall w^i} \sum_{\forall w^j} f(X_{w^i, w^j}) (s(w^i, w^j) - \log(X_{w^i, w^j})) \quad (3.23)$$

325 Where  $f(x)$  is a weighing between 0 and 1 given by:

$$326 \quad f(x) = \begin{cases} \left(\frac{x}{100}\right)^{0.75} & x < 100 \\ 1 & \text{otherwise} \end{cases} \quad (3.24)$$

327 This can be considered as a saturating variant of the effective weighing of skip-gram  
 328 being  $X_{w^i, w^j}$ .

329 While GloVe out-performed skip-gram in initial tests subsequent more extensive  
 330 testing in Levy et al. (2015) with more tuned parameters, found that skip-gram  
 331 marginally out-performed GloVe on all tasks.

### LW Move to start of section

#### Implementing GloVe

To implement GloVe in any technical programming language with good support for optimisation is quiet easy, as it is formed into a pure optimization problem. It is also easy to do in a neural network framework, as these always

include an optimiser. Though unlike in normal neural network training there are no discrete training cases, just the global co-occurrence statistics.

### 332 3.3.2 Further Equivalence of Co-location Prediction to Factorisation

333 GloVe highlights the relationship between the co-located word prediction neural  
334 network models, and the more traditional non-negative matrix factorization of co-  
335 location counts used in topic modeling. Very similar properties were also explored  
336 for skip-grams with negative sampling in Levy and Yoav (2014) and in Li et al. (2015)  
337 with more direct mathematical equivalence to weighed co-occurrence matrix factori-  
338 sation; Later, Cotterell et al. (2017) showed the equivalence to exponential principal  
339 component analysis (PCA). Landgraf and Jeremy (2017) goes on to extend this to  
340 show that it is a weighted logistic PCA, which is a special case of the exponential  
341 PCA. Many works exist in this area now.

### 342 3.3.3 Conclusion

343 We have now concluded that neural predictive co-location models are functionally  
344 very similar to matrix factorisation of co-location counts with suitable weightings,  
345 and suitable similarity metrics. One might now suggest a variety of word embeddings  
346 to be created from a variety of different matrix factorisations with different weightings  
347 and constraints. Traditionally large matrix factorisations have significant problems  
348 in terms of computational time and memory usage. A common solution to this, in  
349 applied mathematics, is to handle the factorisation using an iterative optimisation  
350 procedure. Training a neural network, such as skip-gram, is indeed just such an  
351 iterative optimisation procedure.

## 352 3.4 Hierarchical Softmax and Negative Sampling

353 Hierarchical softmax, and negative sampling are effectively alternative output layers  
354 which are computationally cheaper to evaluate than regular softmax. They are power-  
355 ful methods which pragmatically allow for large speed-up in any task which involves  
356 outputting very large classification probabilities – such as language modelling.

### 357 3.4.1 Hierarchical Softmax

358 Hierarchical softmax was first presented in Morin and Bengio (2005). Its recent use  
359 was popularised by Mikolov et al. (2013a), where words are placed as leaves in a  
360 Huffman tree, with their depth determined by their frequency.



### SemHuff

It can be noted that the Huffman encoding scheme specifies only the depth of a given word in the tree. It does not specify the order. Yang et al. (2016) make use of the BlossomV algorithm (Kolmogorov 2009) to pair the nodes on each layer according to their similarity. They found that on the language modelling task this improved performance, in the way one would expect. They used a lexical resource to determine similarity, however noted that a prior trained word-embedding model could be used to define similarity instead – the new encoding can then be used to define a new model which will find new (hopefully better) embeddings. This is similar to the original method used by (Morin and Bengio 2005), but only using the similarity measure for reordering nodes at the same depth, after the depth is decided by Huffman encoding. In our own experimentation, when applying it to other tasks, we did not see large improvements. It is nevertheless a very interesting idea, and quite fun to implement and observe the results.

One of the most expensive parts of training and using a neural language model is to calculate the final softmax layer output. This is because the softmax denominator includes terms for each word in the vocabulary. Even if only one word's probability is to be calculated, one denominator term per word in the vocabulary must be evaluated. In hierarchical softmax, each word (output choice), is considered as a leaf on a binary tree. Each level of the tree roughly halves the space of the output words to be considered. The final level to be evaluated for a given word contains the word's leaf-node and another branch, which may be a leaf-node for another word, or a deeper sub-tree

The tree is normally a Huffman tree (Huffman 1952), as was found to be effective by Mikolov et al. (2013a). This means that for each word  $w^i$ , the word's depth (i.e its code's length)  $l(w^i)$  is such that over all words:  $\sum_{\forall w^j \in \mathbb{V}} P(w^j) \times l(w^j)$  is minimised. Where  $P(w^i)$  is word  $w^i$ 's unigram probability, and  $\mathbb{V}$  is the vocabulary. The approximate solution to this is that  $l(w^i) \approx -\log_2(P(w^i))$ . From the tree, each word can be assigned a code in the usual way, with 0 for example representing taking one branch, and 1 representing the other. Each point in the code corresponds to a node in the binary tree, which has decision tied to it. This code is used to transform the large multinomial softmax classification into a series of binary logistic classifications. It is important to understand that the layers in the tree are not layers of the neural network in the normal sense – the layers of the tree do not have an output that is used as the input to another. The layers of the tree are rather subsets of the neurons on the output layer, with a relationship imparted on them.

It was noted by Mikolov et al. (2013a), that for vocabulary  $\mathbb{V}$ :

- Using normal softmax would require each evaluation to perform  $|\mathbb{V}|$  operations.
- Using hierarchical softmax with a balanced tree, would mean the expected number of operations across all words would be  $\log_2(|\mathbb{V}|)$ .
- Using a Huffman tree gives the expected number of operations  $\sum_{\forall w^j \in \mathbb{V}} -P(w^j) \log_2(P(w^i)) = H(\mathbb{V})$ , where  $H(\mathbb{V})$  is the unigram entropy of words in the training corpus.

**Fig. 3.10** Tree for 2 words

389 The worse case value for the entropy is  $\log_2(|\mathbb{V}|)$ . In-fact Huffman encoding is  
 390 provably optimal in this way. As such this is the minimal number of operations  
 391 required in the average case.

### 3.4.1.1 An Incredibly Gentle Introduction to Hierarchical Softmax

392 In this section, for brevity, we will ignore the bias component of each decision at  
 393 each node. It can either be handled nearly identically to the weight; or the matrix  
 394 can be written in *design matrix form* with an implicitly appended column of ones;  
 395 or it can even be ignored in the implementation (as was done in Mikolov et al.  
 396 2013a). The reasoning for being able to ignore it is that the bias in normal softmax  
 397 encodes unigram probability information; in hierarchical softmax, when used with  
 398 the common Huffman encoding, its the tree's depth in tree encodes its unigram  
 399 probability. In this case, not using a bias would at most cause an error proportionate  
 400 to  $2^{-k}$ , where  $k$  is the smallest integer such that  $2^{-k} > P(w^i)$ .

#### 401 First Consider a Binary Tree with just 1 Layer and 2 Leaves

402 The leaves are  $n^{00}$  and  $n^{01}$ , each of these leaf nodes corresponds to a word from the  
 403 vocabulary, which has size two, for this toy example. **This is shown in Fig 3.10.**

404 From the initial root which we call  $n^0$ , we can go to either node  $n^{00}$  or node  $n^{01}$ ,  
 405 based on the input from the layer below which we will call  $\tilde{z}$  (Fig. 3.10).

406 Here we write  $n^{01}$  to represent the event of the first non-root node being the branch  
 407 given by following left branch, while  $n^{01}$  being to follow the right branch. (The order  
 408 within the same level is arbitrary in any-case, but for our visualisation purposes we'll  
 409 used this convention.)

410 We are naming the root node as a notation convenience so we can talk about the  
 411 decision made at  $n^0$ . Note that  $P(n^0) = 1$ , as all words include the root-node on their  
 412 path.

413 We wish to know the probability of the next node being the left node (i.e.  $P(n^{00} |$   
 414  $\tilde{z})$ ) or the right-node (i.e.  $P(n^{01} | \tilde{z})$ ). As these are leaf nodes, the prediction either  
 415 equivalent to the prediction of one or the other of the two words in our vocabulary.

416 We could represent the decision with a softmax with two outputs. However, since  
 417 it is a binary decision, we do not need a softmax, we can just use a sigmoid.

$$419 \quad P(n^{01} | \tilde{z}) = 1 - P(n^{00} | \tilde{z}) \quad (3.25)$$

420 The weight matrix for a sigmoid layer has a number of columns governed by the  
 421 number of outputs. As there is only one output, it is just a row vector. We are going  
 422 to index it out of a matrix  $V$ . For the notation, we will use index 0 as it is associated

423 with the decision at node  $n^0$ . Thus we call it  $V_{0,:}$ .

424  $P(n^{00} | \tilde{z}) = \sigma(V_{0,:}\tilde{z})$  (3.26)

425  $P(n^{01} | \tilde{z}) = 1 - \sigma(V_{0,:}\tilde{z})$  (3.27)

427 Note that for the sigmoid function:  $1 - \sigma(x) = \sigma(-x)$ . Allowing the formulation to  
428 be written:

429  $P(n^{01} | \tilde{z}) = \sigma(-V_{0,:}\tilde{z})$  (3.28)

430 thus

431  $P(n^{0i} | \tilde{z}) = \sigma((-1)^i V_{0,:}\tilde{z})$  (3.29)

432 Noting that in Eq. (3.29),  $i$  is either 0 (with  $-1^0 = 1$ ) or 1 (with  $-1^1 = -1$ ).

### $V_{0,:}\tilde{z}$ is a dot product

We mentioned in the marginalia earlier, but just as an extra reminder: the matrix product of a row vector like  $V_{0,:}$  with a (column) vector like  $\tilde{z}$  is their vector dot product.

### Now Consider 2 Layers with 3 Leaves

433 Consider a tree with nodes:  $n^0, n^{00}, n^{000}, n^{001}, n^{01}$ . The leaves are  $n^{000}, n^{001}$ , and  
434  $n^{01}$ , each of which represents one of the 3 words from the vocabulary (Fig. 3.11).

435 From earlier we still have:

436  $P(n^{00} | \tilde{z}) = \sigma(V_{0,:}\tilde{z})$  (3.30)

437  $P(n^{01} | \tilde{z}) = \sigma(-V_{0,:}\tilde{z})$  (3.31)

440 We must now to calculate  $P(n^{000} | \tilde{z})$ . Another binary decision must be made at  
441 node  $n^{00}$ . The decision at  $n^{00}$  is to find out if the predicted next node is  $n^{000}$  or  $n^{001}$ .  
442 This decision is made, with the assumption that we have reached  $n^{00}$  already.

443 So the decision is defined by  $P(n^{000} | z, n^{00})$  is given by:

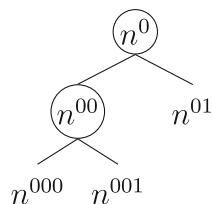
444  $P(n^{000} | \tilde{z}) = P(n^{000} | \tilde{z}, n^{00}) P(n^{00} | \tilde{z})$  (3.32)

445  $P(n^{000} | \tilde{z}, n^{00}) = \sigma(V_{00,:}\tilde{z})$  (3.33)

446  $P(n^{001} | \tilde{z}, n^{00}) = \sigma(-V_{00,:}\tilde{z})$  (3.34)

Fig. 3.11 Tree for 3 words

LW: close gap by moving caption or image



448 We can use the conditional probability chain rule to recombine to compute the three  
 449 leaf nodes final probabilities.

$$450 \quad P(n^{01} | \tilde{z}) = \sigma(-V_{0,:}\tilde{z}) \quad (3.35)$$

$$451 \quad P(n^{000} | \tilde{z}) = \sigma(V_{00,:}\tilde{z})\sigma(V_{0,:}\tilde{z}) \quad (3.36)$$

$$452 \quad P(n^{001} | \tilde{z}) = \sigma(-V_{00,:}\tilde{z})\sigma(V_{0,:}\tilde{z}) \quad (3.37)$$

#### 454 Continuing this Logic

455 Using this system, we know that for a node encoded at position  $[0t^1t^2t^3\dots t^L]$ , e.g.  
 456 [010\dots 1], its probability can be found recursively as:

$$457 \quad P(n^{0t^1\dots t^L} | \tilde{z}) = \underbrace{P(n^{0t^1\dots t^L} | \tilde{z}, n^{0t^1\dots t^{L-1}})}_{458} P(n^{0t^1\dots t^{L-1}} | \tilde{z}) \quad (3.38)$$

460 Thus:

$$461 \quad P(n^{0t^1} | \tilde{z}) = \sigma((-1)^{t^1} V_{0,:}\tilde{z}) \quad (3.39)$$

$$462 \quad P(n^{0t^1,t^2} | \tilde{z}, n^{0t^1}) = \sigma((-1)^{t^2} V_{0t^1,:}\tilde{z}) \quad (3.40)$$

$$463 \quad P(n^{0t^1\dots t^i} | \tilde{z}, n^{0t^1\dots t^{i-1}}) = \sigma((-1)^{t^i} V_{0t^1\dots t^{i-1},:\tilde{z}}) \quad (3.41)$$

465 The conditional probability chain rule, is applied to get:

$$466 \quad P(n^{0t^1\dots t^L} | \tilde{z}) = \prod_{i=1}^{i=L} \sigma((-1)^{t^i} V_{0t^1\dots t^{i-1},:\tilde{z}}) \quad (3.42)$$

#### 467 3.4.1.2 Formulation

##### Combining multiplications

If one wants to find both  $V_{00,:}\tilde{z}$  and  $V_{0,:}\tilde{z}$ , then this can be done using matrices simultaneously, thus potentially taking advantage of optimized matrix multiplication routines.

$$\begin{bmatrix} V_{0,:} \\ V_{00,:} \end{bmatrix} z = \begin{bmatrix} V_{0,:}\tilde{z} \\ V_{00,:}\tilde{z} \end{bmatrix}$$

Thus the whole product for all of the decisions can be written as  $V\tilde{z}$ . The problem then becomes indexing the relevant node rows.

However computing every single decision is beyond what is required for most uses: hierarchical softmax lets us only compute the decisions that are on the path to the word-leaf we which we wish to query. Computing all of them is beyond what is required.



Packing tree node elements into a matrix with fast indexing is a non-trivial problem. The details on optimising such multiplications and tree packing are beyond the scope of this book.

In general there may be very little scope here for optimisation, as on most hardware (and BLAS systems) matrix products with  $n$  columns, takes a similar amount of time to  $n$  vector dot products. As such storing the row vectors of  $V$  in a hashmap indexed by node-path, and looping over them as required may be more practical.

In languages/libraries with slow looping constructs (numpy, R, octave), where calling into suitable library routines is much faster, this may give some speed-up; but even here it is likely to be minor. The time may be better spent writing a C extension library to do this part of the program. Or learning to use a language with fast for loops (e.g. Julia 2014).

468 The formulation above is not the same as in other works. This subsection shows the  
 469 final steps to reach the conventional form used in Mikolov et al. (2013b).

470 Here we have determined that the 0th/left branch represents the positive choice,  
 471 and the other probability is defined in terms of this. It is equivalent to have the  
 472 1th/right branch representing the positive choice:

$$473 \quad P(n^{0t^1 \dots t^L} | \tilde{z}) = \prod_{i=1}^{i=L} \sigma\left((-1)^{t^i+1} V_{0t^1 \dots t^{i-1}, :} \tilde{z}\right) \quad (3.43)$$

474 or to allow it to vary per node: as in the formulation of Mikolov et al. (2013b). In that  
 475 work they use  $ch(n)$  to represent an arbitrary child node of the node  $n$  and use an  
 476 indicator function  $\llbracket a = b \rrbracket = \begin{cases} 1 & a = b \\ -1 & a \neq b \end{cases}$  such that they can write  $\llbracket n^b = ch(n^a) \rrbracket$   
 477 which will be 1 if  $n^a$  is an arbitrary (but consistent) child of  $n^b$ , and 0 otherwise.

$$478 \quad P(n^{0t^1 \dots t^L} | \tilde{z}) = \\ 479 \quad \prod_{i=1}^{i=L} \sigma\left(\llbracket n^{0t^1 \dots t^i} = ch(n^{0t^1 \dots t^{i-1}}) \rrbracket V_{0t^1 \dots t^{i-1}, :} \tilde{z}\right) \quad (3.44)$$

481 There is no functional difference between the three formulations. Though the final  
 482 one is perhaps a key reason for the difficulties in understanding the hierarchical  
 483 softmax algorithm.

### How does this relate to word vectors?

After the length of this section, one may have forgotten why we are doing this in the first place. Recall that CBOW, skip-gram and all other language modelling based word embedding methods are based around predicting  $P(w^o | w^i, \dots, w^j)$  for some words. For skip-gram that is just  $P(w^o | w^i)$ .

LW: move this if possible to avoid splitting box over 2 pages

The term  $n^{0t^1 \dots t^L}$  in  $P(n^{0t^1 \dots t^L} | z)$ , just represents as a path through the tree to the leaf node which represents the word  $w^o$ . i.e  $P(n^{0t^1 \dots t^L} | z) = P(w^o | z)$ . The output of the final hidden layer is  $z$  (i.e. the  $z$  is the input to the output layer) In normal language models  $z$  encodes all the information about what the model knows of predictions based on  $w^i, \dots, w^j$ .  $z$  is thus a proxy term in the conditional probability for those words. In skip-gram there is no hidden layer, and it is just  $z = C_{:,w^i}$  proxying only for  $w_i$ , and the model defines its probability output by  $P(w^o | w^i) = P(w^o | C_{:,w^i})$ .

#### 484 3.4.1.3 Loss Function

##### The gradient calculations

They are not fun. They never are for back-propagation. We recommend using a framework with automated differentiation, and/or performing gradient checks against a numerical differentiation tool (simple finite-differencing will do in a pinch).

485 Using normal softmax, during the training, the cross-entropy between the model's  
 486 predictions and the ground truth as given in the training set is minimised. Cross  
 487 entropy is given by

$$488 CE(P^*, P) = \sum_{\forall w^i \in \mathbb{V}} \sum_{\forall z^j \in \mathbb{Z}} -P^*(w^i | z^j) \log P(w^i | z^j) \quad (3.45)$$

489 where  $P^*$  is the true distribution, and  $P$  is the approximate distribution given by  
 490 our model (in other sections we have abused notation to use  $P$  for both).  $\mathbb{Z}$  is the set  
 491 of values that are input into the model, (or equivalently the values derived from  
 492 them from lower layers) – the context words in language modelling.  $\mathbb{V}$  is the set  
 493 of outputs, the vocabulary in language modeling. The training dataset  $\mathcal{X}$  consists of  
 494 pairs from  $\mathbb{V} \times \mathbb{Z}$ .

495 The true probabilities (from  $P^*$ ) are implicitly given by the frequency of the  
 496 training pairs in the training dataset  $\mathcal{X}$ .

$$497 Loss = CE(P^*, P) = \frac{1}{|\mathcal{X}|} \sum_{\forall (w^i, z^i) \in \mathcal{X}} -\log P(w^i | z^i) \quad (3.46)$$

498 The intuitive understanding of this, is that we are maximising the probability estimate  
 499 of all pairings which actually occur in the training set, proportionate to how often  
 500 they occur. Note that the  $\mathbb{Z}$  can be non-discrete values, as was the whole benefit of  
 501 using embeddings, as discussed in Sect. 3.1.1.

502 This works identically for hierarchical softmax as for normal softmax. It is simply  
 503 a matter of substituting in the (different) equations for  $P$ . Then applying back-  
 504 propagation as usual.

### 505 3.4.2 Negative Sampling

506 Negative sampling was introduced in Mikolov et al. (2013b) as another method to  
 507 speed up this problem. Much like hierarchical softmax in its purpose. However,  
 508 negative sampling does not modify the network's output, but rather the loss function.

509 Negative Sampling is a simplification of Noise Contrast Estimation (Gutmann and  
 510 Hyvärinen 2012). Unlike Noise Contrast Estimation (and unlike softmax), it does  
 511 not in fact result in the model converging to the same output as if it were trained with  
 512 softmax and cross-entropy loss. However the goal with these word embeddings is not  
 513 to actually perform the language modelling task, but only to capture a high-quality  
 514 vector representation of the words involved.

#### 515 3.4.2.1 A Motivation of Negative Sampling

516 Recall from Sect. 3.2.2 that the (supposed) goal, is to estimate  $P(w^j | w^i)$ . In truth,  
 517 the goal is just to get a good representation, but that is achieved via optimising  
 518 the model to predict the words. In Sect. 3.2.2 we considered the representation of  
 519  $P(w^j | w^i)$  as the  $w^j$ th element of the softmax output.

$$520 P(w^j | w^i) = \text{smax}(V C_{:,w^i})_{w^j} \quad (3.47)$$

$$521 P(w^j | w^i) = \frac{\exp(V_{w^j,:} C_{:,w^i})}{\sum_{k=1}^N \exp(V_{k,:} C_{:,k})} \quad (3.48)$$

#### Why is not using softmax wrong?

The notation abuse may be hiding just how bad it is to not use softmax.  
 Recall that the true meaning of  $P(w^j | w^i)$  is actually  $P(W^j=w^j | W^i=w^i)$ .  
 By not using softmax, with its normalising denominator this means that:  
 $\sum_{\forall w^j \in \mathbb{V}} P(w^j | w^i) \neq 1$  (except by coincidence).

523 This is not the only valid representation. One could use a sigmoid neuron for a  
 524 direct answer to the co-location probability of  $w^j$  occurring near  $w^i$ . Though this  
 525 would throw away the promise of the probability distribution to sum to one across all  
 526 possible words that could be co-located with  $w^i$ . That promise could be enforced by  
 527 other constraints during training, but in this case it will not be. It is a valid probability  
 528 if one does not consider it as a single categorical prediction, but rather as independent  
 529 predictions.

$$530 P(w^j | w^i) = \sigma(V C_{:,w^i})_{w^j} \quad (3.49)$$

$$531 \text{i.e. } P(w^j | w^i) = \sigma(V_{w^j,:} C_{:,w^i}) \quad (3.50)$$

533 Lets start from the cross-entropy loss. In training word  $w^j$  does occur near  $w^i$ , we  
 534 know this because they are a training pair presented from the training dataset  $\mathcal{X}$ .

535 Therefore, since it occurs, we could make a loss function based on minimising the  
 536 negative log-likelihood of all observations.

$$537 \quad Loss = \sum_{\forall (w^i, w^j) \in \mathcal{X}} -\log P(w^j | w^i) \quad (3.51)$$

538 This is the cross-entropy loss, excluding the scaling factor for how often it occurs.

### **Loss Function**

Readers may want to reread Sect. 3.4.1.3 to brush up on how we use the training dataset as a ground truth probability estimate implicitly when using cross-entropy loss. When doing so one should remember that that the conditioning term,  $z$ , for skip-grams is the co-located words as there is no hidden layer.

539 However, we are not using softmax in the model output, which means that there  
 540 is no trade off for increasing (for example)  $P(w^1 | w^i)$  versus  $P(w^2 | w^i)$ . This  
 541 thus admits the trivially optimal solution  $\forall w^j \in \mathbb{V} P(w^j | w^i) = 1$ . This is obvi-  
 542 ously wrong – even beyond not being a proper distribution – some words are more  
 543 commonly co-occurring than others.

544 So from this we can improve the statement. What is desired from the loss function  
 545 is to reward models that predict the probability of words that *do* co-occur as being  
 546 higher, than the probability of words that *do not*. We know that  $w^j$  does occur  
 547 near  $w^i$  as it is in the training set. Now, let us select via some arbitrary means a  
 548  $w^k$  that does not – a negative sample. We want the loss function to be such that  
 549  $P(w^k | w^i) < P(w^j | w^i)$ . So for this single term in the loss we would have:

$$550 \quad loss(w^j, w^i) = \log P(w^k | w^i) - \log P(w^j | w^i) \quad (3.52)$$

### **Most words do not co-occur**

Some simple reasoning can account for this as a reasonable consequence of Zipf's law (Zipf 1949) and a prior of the principle of indifference, but there is a further depth to it as explained by Ha et al. (2009).

551 The question is then: how is the negative sample  $w^k$  to be found? One option would  
 552 be to deterministically search the corpus for these negative samples, making sure to  
 553 never select words that actually do co-occur. However that would require enumerating  
 554 the entire corpus. We can instead just pick them randomly, we can sample from the  
 555 unigram distribution. As statistically, in any given corpus most words do not co-  
 556 occur, a randomly selected word in all likelihood will not be one that truly does  
 557 co-occur – and if it is, then that small mistake will vanish as noise in the training,  
 558 overcome by all the correct truly negative samples.

### Is Eq. (3.53) a function?

No, at the point at which the Loss started including randomly selected samples, it ceased to be a function in the usual mathematical sense. It is still a function in the common computer programming sense though – it is just not deterministic.

At this point, we can question, why limit ourselves to one negative sample? We could take many, and do several at a time, and get more confidence that  $P(w^j | w^i)$  is indeed greater than other (non-existent) co-occurrence probabilities. This gives the improved loss function of

$$\text{loss}(w^j, w^i) = \left( \sum_{\forall w^k \in \text{samples}(D^{1g})} \log P(w^k | w^i) \right) - \log P(w^j | w^i) \quad (3.53)$$

where  $D^{1g}$  stands for the unigram distribution of the vocabulary and  $\text{samples}(D^{1g})$  is a function that returns some number of samples from it.

Consider, though is this fair to the samples? We are taking them as representatives of all words that do not co-occur. Should a word that is unlikely to occur at all, *but was unlucky enough to be sampled*, contribute the same to the loss as a word that was very likely to occur? More reasonable is that the loss contribution should be in proportion to how likely the samples were to occur. Otherwise it will add unexpected changes and result in noisy training. Adding a weighting based on the unigram probability ( $P^{1g}(w^k)$ ) gives:

$$\text{loss}(w^j, w^i) = \underbrace{\left( \sum_{\forall w^k \in \text{samples}(D^{1g})} P^{1g}(w^k) \log P(w^k | w^i) \right)}_{\text{loss per sample}} - \log P(w^j | w^i) \quad (3.54)$$

The expected value is defined by

$$\mathbb{E}_{X \sim D}[f(x)] = \sum_{\forall x \text{ values for } X} P^d f(x) \quad (3.55)$$

In an abuse of notation, we apply this to the samples, as a sample expected value and write:

$$\sum_{k=1}^{k=n} \mathbb{E}_{w^k \sim D^{1g}} [\log P(w^k | w^i)] \quad (3.56)$$

to be the sum of the  $n$  samples expected values. This notation (abuse) is as used in Mikolov et al. (2013b). It gives the form:

$$\text{loss}(w^j, w^i) = \underbrace{\left( \sum_{k=1}^{k=n} \mathbb{E}_{w^k \sim D^{1g}} [\log P(w^k | w^i)] \right)}_{\text{expected loss per sample}} - \log P(w^j | w^i) \quad (3.57)$$

Consider that the choice of unigram distribution for the negative samples is not the only choice. For example, we might wish to increase the relative occurrence of rare words in the negative samples, to help them fit better from limited training data. This is commonly done via subsampling in the positive samples (i.e. the training cases)). So we replace  $D^{1g}$  with  $D^{ns}$  being the distribution of negative samples from the vocabulary, to be specified as a hyper-parameter of training.

Mikolov et al. (2013b) uses a distribution such that

$$P^{D^{ns}}(w^k) = \frac{P^{D^{1g}}(w^k)^{\frac{2}{3}}}{\sum_{\forall w^o \in \mathbb{V}} P^{D^{1g}}(w^o)^{\frac{2}{3}}} \quad (3.58)$$

which they find to give better performance than the unigram or uniform distributions.

Using this, and substituting in the sigmoid for the probabilities, this becomes:

$$\begin{aligned} loss(w^j, w^i) &= \\ &\left( \sum_{k=1}^{k=n} \mathbb{E}_{w^k \sim D^{ns}} [\log \sigma(V_{w^k,:} C_{:,w^i})] \right) - \log \sigma(V_{w^j,:} C_{:,w^i}) \end{aligned} \quad (3.59)$$

By adding a constant we do not change the optimal value. If we add the constant  $-K$ , we can subtract 1 in each sample term.

$$\begin{aligned} loss(w^j, w^i) &= \\ &\left( \sum_{k=1}^{k=n} \mathbb{E}_{w^k \sim D^{ns}} [-1 + \log \sigma(V_{w^k,:} C_{:,w^i})] \right) - \log \sigma(V_{w^j,:} C_{:,w^i}) \end{aligned} \quad (3.60)$$

Finally we make use of the identity  $1 - \sigma(\tilde{z}) = \sigma(-\tilde{z})$  giving:

$$\begin{aligned} loss(w^j, w^i) &= \\ &- \log \sigma(V_{w^j,:} C_{:,w^i}) - \sum_{k=1}^{k=n} \mathbb{E}_{w^k \sim D^{ns}} [\log \sigma(-V_{w^k,:} C_{:,w^i})] \end{aligned} \quad (3.61)$$

Calculating the total loss over the training set  $\mathcal{X}$ :

$$\begin{aligned} Loss &= - \sum_{\forall (w^i, w^j) \in \mathcal{X}} \\ &\left( \log \sigma(V_{w^j,:} C_{:,w^i}) + \sum_{k=1}^{k=n} \mathbb{E}_{w^k \sim D^{ns}} [\log \sigma(-V_{w^k,:} C_{:,w^i})] \right) \end{aligned} \quad (3.62)$$

612 This is the negative sampling loss function used in Mikolov et al. (2013b). Perhaps  
 613 the most confusing part of this is the notation. Without the abuses around expected  
 614 value, this is written:

$$615 \quad Loss = - \sum_{\forall (w^i, w^j) \in \mathcal{X}} \\ 616 \quad \left( \log \sigma(V_{w^j,:} C_{:,w^i}) + \sum_{\substack{\forall w^k \in \text{samples}(D^{\text{ns}})}} P^{\text{D}^{\text{ns}}}(w^k) \log \sigma(-V_{w^k,:} C_{:,w^i}) \right) \quad (3.63)$$

LW: A lot of the math in this section has extra linebreaks.

I believe this is due to there now being a bit of extra space; due to the removal of the marginalia/side-boxes

## 618 3.5 Natural Language Applications – Beyond Language 619 Modeling

620 While statistical language models are useful, they are of-course in no way the be-all  
 621 and end-all of natural language processing. Simultaneously with the developments  
 622 around representations for the language modelling tasks, work was being done on  
 623 solving other NLP problems using similar techniques (Collobert and Jason 2008).

### 624 3.5.1 Using Word Embeddings as Features

#### Pretrained Word-Embeddings

Pretrained Word Embeddings are available for most models discussed here. They are trained on a lot more data than most people have reasonable access to. It can be useful to substitute word embeddings as a representation in most systems, or to use them as initial value for neural network systems that will learn them as they train the system as a whole. There are many many online pretrained word embeddings. One of the more recent and comprehensive set is that of Bojanowski et al. (2016) (based on a skip-gram extension), <https://fasttext.cc/docs/en/pretrained-vectors.html> They provide embeddings for 294 languages, trained on Wikipedia based on the work of which is an extension to skip-grams.

625 Turian et al. (2010) discuss what is now perhaps the most important use of word  
 626 embeddings. The use of the embeddings as features, in unrelated feature driven  
 627 models. One can find word embeddings using any of the methods discussed above.  
 628 These embeddings can be then used as features instead of, for example bag of words or  
 629 hand-crafted feature sets. Turian et al. (2010) found improvements on the state of the  
 630 art for chunking and Named Entity Recognition (NER), using the word embedding  
 631 methods of that time. Since then, these results have been superseded again using  
 632 newer methods.

## 633 3.6 Aligning Vector Spaces Across Languages

634 Given two vocabulary vector spaces, for example one for German and one for English,  
 635 a natural and common question is if they can be aligned such that one has a single  
 636 vector space for both. Using canonical correlation analysis (CCA) one can do exactly  
 637 that. There also exists generalised CCA for any number of vector spaces (Fu et al.  
 638 2016), as well as kernel CCA for a non-linear alignment.

639 The inputs to CCA, are two sets of vectors, normally expressed as matrices. We  
 640 will call these:  $C \in \mathbb{R}^{n^C \times m^C}$  and  $V \in \mathbb{R}^{n^V \times m^V}$ . They are both sets of vector represen-  
 641 tations, not necessarily of the same dimensionality. They could be the output of any  
 642 of the embedding models discussed earlier, or even a sparse (non-embedding) represen-  
 643 tations such as the point-wise mutual information of the co-occurrence counts.  
 644 The other input is series pairs of elements from within those sets that are to  
 645 be aligned. We will call the elements from that series of pairs from the original sets  
 646  $C^*$  and  $V^*$  respectively.  $C^*$  and  $V^*$  are subsets of the original sets, with the same  
 647 number of representations. In the example of applying this to translation, if each  
 648 vector was a word embedding:  $C^*$  and  $V^*$  would contain only words with a single  
 649 known best translation, and this does not have to be the whole vocabulary of either  
 650 language.

651 By performing CCA one solves to find a series of vectors (also expressed as a  
 652 matrix),  $S = [\tilde{s}^1 \dots \tilde{s}^d]$  and  $T = [\tilde{t}^1 \dots \tilde{t}^d]$ , such that the correlation between  $C^* \tilde{s}^i$   
 653 and  $V^* \tilde{t}^i$  is maximised, with the constraint that for all  $j < i$  that  $C^* \tilde{s}^i$  is uncorrelated  
 654 with  $C^* \tilde{s}^j$  and that  $V^* \tilde{t}^i$  is uncorrelated with  $V^* \tilde{t}^j$ . This is very similar to principal  
 655 component analysis (PCA), and like PCA the number of components to use ( $d$ )  
 656 is a variable which can be decreased to achieve dimensionality reduction. When  
 657 complete, taking  $S$  and  $T$  as matrices gives projection matrices which project  $C$   
 658 and  $V$  to a space where aligned elements are as correlated as possible. The new  
 659 common vector space embeddings are given by:  $CS$  and  $VT$ . Even for sparse inputs  
 660 the outputs will be dense embeddings.

661 Faruqui and Chris (2014) investigated this primarily as a means to use additional  
 662 data to improve performance on monolingual tasks. In this, they found a small and  
 663 inconsistent improvement. However, we suggest it is much more interesting as a  
 664 multi-lingual tool. It allows similarity measures to be made between words of differ-  
 665 ent languages. Gujral et al. (2016) use this as part of a hybrid system to translate out  
 666 of vocabulary words. Klein et al. (2015) use it to link word-embeddings with image  
 667 embeddings.

668 Dhillon et al. (2011) investigated using this to create word-embeddings. We noted  
 669 in Eq. (3.16), that skip-gram maximise the similarity of the output and input embed-  
 670 dings according to the dot-product. CCA also maximises similarity (according the  
 671 correlation), between the vectors from one set, and the vectors for another. As such  
 672 given representations for two words from the same context, initialised randomly,  
 673 CCA could be used repeatedly to optimise towards good word embedding capturing  
 674 shared meaning from contexts. This principle was used by Dhillon et al. (2011),  
 675 though their final process more complex than described here. It is perhaps one of the

more unusual ways to create word embeddings as compared to any of the methods discussed earlier.

Aligning embeddings using linear algebra after they are fully trained is not the only means to end up with a common vector space. One can also directly train embeddings on multiple languages concurrently as was done in Shi et al. (2015), amongst others. Similarly, on the sentence embedding side (Zou et al. 2013), and (Socher et al. 2014) train embeddings from different languages and modalities (respectively) directly to be near to their partners (these are discussed in Chap. 5). A survey paper on such methods was recently published by Ruder (2017).

685

## 686 References

- Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *The Journal of Machine Learning Research*, 137–186.
- Bezanson, Jeff, Alan Edelman, Stefan Karpinski, and Viral B. Shah. 2014. Julia: A fresh approach to numerical computing. [arXiv:1411.1607](https://arxiv.org/abs/1411.1607) [cs.MS]
- Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *The Journal of Machine Learning Research* 3: 993–1022.
- Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. [arXiv:1607.04606](https://arxiv.org/abs/1607.04606).
- Bolukbasi, Tolga, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems*, 4349–4357.
- Brown, Peter F., Peter V. Desouza, Robert L. Mercer, Vincent J. Della, and Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Computational Linguistics* 18 (4): 467–479.
- Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356 (6334), 183–186. ISSN: 0036-8075, <https://doi.org/10.1126/science.aal4230>, <http://science.scienmag.org/content/356/6334/183.full.pdf>.
- Collobert, Ronan and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, 160–167. ACM.
- Cotterell, Ryan, Adam Poliak, Benjamin Van Durme, and Jason Eisner. 2017. Explaining and generalizing skip-gram through exponential family principal component analysis. In *EACL 2017*, 175.
- Dhillon, Paramveer, Dean P. Foster, and Lyle H. Ungar. 2011. Multi-view learning of word embeddings via CCA. In *Advances in Neural Information Processing Systems*, 199–207.
- Dumais, Susan T., George W. Furnas, Thomas K. Landauer, Scott Deerwester, and Richard Harshman. 1988. Using latent semantic analysis to improve access to textual information. In *Proceedings of the SIGCHI conference on human factors in computing systems*, 281–285. ACM.
- Faruqui, Manaal and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Association for computational linguistics*.
- Fu, X., K. Huang, E. E. Papalexakis, H. A. Song, P. P. Talukdar, N. D. Sidiropoulos, C. Faloutsos, and T. Mitchell. 2016. Efficient and distributed algorithms for large-scale generalized canonical

- correlations analysis. In *2016 IEEE 16th international conference on data mining (ICDM)*, 871–876. <https://doi.org/10.1109/ICDM.2016.0105>.
- Gladkova, Anna, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: What works and what doesn't. In *SRW@ HLT-NAACL*, 8–15.
- Goodman, Joshua. 2001. A bit of progress in language modeling. In *CoRR*. arXiv:cs.CL/0108005.
- Gujral, Biman, Huda Khayrallah, and Philipp Koehn. 2016. Translation of unknown words in low resource languages. In *Proceedings of the conference of the association for machine translation in the Americas (AMTA)*.
- Gutmann, Michael U. and Aapo Hyvärinen. 2012. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research* 13: 307–361.
- Ha, Le Quan, Philip Hanna, Ji Ming, and F. Jack Smith. 2009. Extending Zipf's law to n-grams for large corpora. *Artificial Intelligence Review* 32 (1): 101–113.
- Huffman, DavidA. 1952. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE* 40 (9): 1098–1101.
- Jozefowicz, Rafal, Wojciech Zaremba, and Ilya Sutskever. 2015. An empirical exploration of recurrent network architectures. In *Proceedings of the 32nd international conference on machine learning (ICML-15)*, 2342–2350.
- Katz, Slava M. 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing* 35 (3): 400–401.
- Klein, Benjamin, Guy Lev, Gil Sadeh, and Lior Wolf. 2015. Associating neural word embeddings with deep image representations using fisher vectors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4437–4446.
- Kneser, Reinhard and Hermann Ney. 1995. Improved backing-off for M-gram language modeling. In *1995 international conference on acoustics, speech, and signal processing, 1995. ICASSP-95*, vol. 1, 181–184. IEEE.
- Kolmogorov, Vladimir. 2009. Blossom V: A new implementation of a minimum cost perfect matching algorithm. *Mathematical Programming Computation* 1 (1): 43–67.
- Landgraf, Andrew J. and Jeremy Bellay. 2017. Word2vec skip-gram with negative sampling is a weighted logistic PCA. In *CoRR*. arXiv:1705.09755.
- Levy, Omer and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems*, 2177–2185.
- Levy, Omer, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics* 3: 211–225. ISSN: 2307-387X.
- Li, Yitan, Linli Xu, Fei Tian, Liang Jiang, Xiaowei Zhong, and Enhong Chen. 2015. Word embedding revisited: A new representation learning and explicit matrix factorization perspective. In *IJCAI*, 3650–3656.
- Lin, Yuri, Jean-Baptiste Michel, Erez Lieberman Aiden, Jon Orwant, Will Brockman, and Slav Petrov. 2012. Syntactic annotations for the google books ngram corpus. In *Proceedings of the ACL 2012 system demonstrations*, 169–174. Association for Computational Linguistics.
- Maaten, Laurens van der and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9: 2579–2605.
- Mikolov, Tomas, Martin Karafiat, Lukas Burget, Jan Cernocky, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. *Interspeech* 2: 3.
- Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. arXiv:1301.3781.

- Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.
- Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, 746–751.
- Morin, Frederic and Yoshua Bengio. 2005. Hierarchical probabilistic neural network language model. In *Proceedings of the international workshop on artificial intelligence and statistics (Citeseer)*, 246–252.
- Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP 2014)*, 1532–1543.
- Rosenfeld, Ronald. 2000. Two decades of statistical language modeling: Where do we go from here? *Proceedings of the IEEE* 88 (8): 1270–1278. <https://doi.org/10.1109/5.880083>.
- Ruder, Sebastian. 2017. A survey of cross-lingual embedding models. In *CoRR*. arXiv:1706.04902.
- Schwenk, Holger. 2004. Efficient training of large neural networks for language modeling. In *2004. Proceedings. 2004 IEEE international joint conference on neural networks*, vol. 4, 3059–3064. IEEE.
- Shi, Tianze, Zhiyuan Liu, Yang Liu, and Maosong Sun. 2015. Learning cross-lingual word embeddings via matrix co-factorization. In *ACL* (2), 567–572.
- Socher, Richard, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, and Y.Ng Andrew. 2014. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics* 2: 207–218.
- Sundermeyer, Martin, Ralf Schlüter, and Hermann Ney. 2012. LSTM neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association*.
- Tomas, Mikolov. 2012. Statistical language models based on neural networks. Ph.D. thesis, Brno University of Technology.
- Turian, Joseph, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, 384–394. Association for Computational Linguistics.
- Yang, Zhixuan, Chong Ruan, Caihua Li, and Junfeng Hu. 2016. Optimize hierarchical softmax with word similarity knowledge. In *17th international conference on intelligent text processing and computational linguistics (CICLing)*.
- Zipf, G.K. 1949. *Human behavior and the principle of least effort: An introduction to human ecology*. Cambridge: Addison-Wesley Press.
- Zou, Will Y., Richard Socher, Daniel M. Cer, and Christopher D. Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *EMNLP*, 1393–1398.

# Author Queries

## Chapter 3

Query Refs.	Details Required	Author's response
AQ1	Please provide high-resolution source file for Fig. 3.1	
AQ2	Please check and confirm if the inserted citation of Figs. 4.7, 4.10, 4.11 are correct. If not, please suggest an alternate citation.	

LW:

AQ1: If I forget to send you a high-res source for figure 3.1 alongside this document, send me an email to remind me to do so.

AQ2: Checked corrections inline if required. (I believe intended was 3.7, 3.10, 3.11)

# Metadata of the chapter that will be visualized in SpringerLink

Book Title	Neural Representations of Natural Language	
Series Title		
Chapter Title	Word Sense Representations	
Copyright Year	2019	
Copyright HolderName	Springer Nature Singapore Pte Ltd.	
Corresponding Author	Family Name Particle Given Name Prefix Suffix Role Division Organization Address Email	<b>White</b> <b>Lyndon</b> <b>Togneri</b> <b>Roberto</b> <b>Liu</b> <b>Wei</b> <b>Bennamoun</b> <b>Mohammed</b>
Author	Family Name Particle Given Name Prefix Suffix Role Division Organization Address Email	School of Electrical, Electronic <b>and</b> Computer Engineering The University of Western Australia Perth, WA, Australia LYNDON.WHITE@RESEARCH.UWA.EDU.AU
Author	Family Name Particle Given Name Prefix Suffix Role Division Organization Address Email	School of Electrical, Electronic <b>and</b> Computer Engineering The University of Western Australia Perth, WA, Australia
Author	Family Name Particle Given Name Prefix Suffix Role Division Organization Address Email	School of Computer Science and Software Engineering The University of Western Australia Perth, WA, Australia
Author	Family Name Particle Given Name Prefix Suffix	

Role  
Division School of Computer Science and Software Engineering  
Organization The University of Western Australia  
Address Perth, WA, Australia  
Email

---

Abstract

In this chapter, techniques for representing the multiple meanings of a single word are discussed. This is a growing area, and is particularly important in languages where polysemous and homonymous words are common. This includes English, but it is even more prevalent in Mandarin for example. The techniques discussed can broadly be classified as lexical word sense representation, and as word sense induction. The inductive techniques can be sub-classified as clustering -based or as prediction-based.

---



# Word Sense Representations

4

LW: Where is the chapter abstract gone?

- 1a. In a literal, exact, or actual sense; not figuratively, allegorically, etc*
- 1b. Used to indicate that the following word or phrase must be taken in its literal sense, usually to add emphasis*
- 1c. colloq. Used to indicate that some (frequently conventional) metaphorical or hyperbolical expression is to be taken in the strongest admissible sense: ‘virtually, as good as’; (also) ‘completely, utterly, absolutely’ ...*
- 2a With reference to a version of something, as a transcription, translation, etc.: in the very words, word for word*
- 2b. In extended use. With exact fidelity of representation; faithfully*
- 3a. With or by the letters (of a word). Obs. rare*
- 3b. In or with regard to letters or literature. Obs. rare*

The seven senses of literally, Oxford English Dictionary, 3rd ed., 2011

- 1 In this chapter, techniques for representing the multiple meanings of a single word are  
2 discussed. This is a growing area, and is particularly important in languages where  
3 polysemous and homonymous words are common. This includes English, but it is  
4 even more prevalent in Mandarin for example. The techniques discussed can broadly  
5 be classified as lexical word sense representation, and as word sense induction. The  
6 inductive techniques can be sub-classified as clustering-based or as prediction-based.



## 4.1 Word Senses

Words have multiple meanings. A single representation for a word cannot truly describe the correct meaning in all contexts. It may have some features that are applicable to some uses but not to others, it may be an average of all features for all uses, or it may only represent the most common sense. For most word-embeddings it will be an unclear combination of all of the above. Word sense embeddings attempt to find representations not of words, but of particular senses of words.

### Polysemous/Homonymous

A word with multiple meanings i.e. senses. For NLP representational purposes polysemous and homonymous are synonymous.

LW: Move marked boxes to somewhere after location marked \*\*\*

### Part of Speech/POS

The syntactic category a word belongs to. Different POS tags come from different tag sets. Can be simple as the WordNet tag set: noun, adjective, verb, etc. or complex as in the Brown tag set: VBG-verb, gerund/present participle, NN-noun, singular or mass.

### Word-use

An occurrence of a word in a text, such as a training corpus. Each word will have multiple uses in a text. Each word-use will only have one particular meaning and will thus belong to one synset.

### Lemma

The base form of the word as defined by a lexicographical resource. It is normally closely related to (often identical to) the *stem* which is the word's root form with all morphological inflections (e.g. tenses) removed.

### Lexeme

The set of words that share a common lemma: go, going, goes, and went all belong to the lexeme headed by the lemma go

### Synset

A synset is a set of synonymous words: that is words that have the same meaning. In lexicographic terms the synset is the core unit of meaning. Identifying

the synset of a word-use is the same as identifying the word sense. Every word sense corresponds to one synset.

### Gloss

A gloss is the dictionary entry for a word sense, it normally includes both the definition and an example of use. In WordNet each synset shares a common gloss.

### Lemmatization LW move to end of section

Lemmatization is the method of converting a word into its lemma. Due to the similarity of the lemma to the stem, this in essence means removing the tense and plurality information (stemming), with some additional special-cases. WordNet is indexed by lemmas, and comes with a lemmatizer called morphy allowing any word to be looked up by lemmatizing it to its lemma

### LW move to end of section

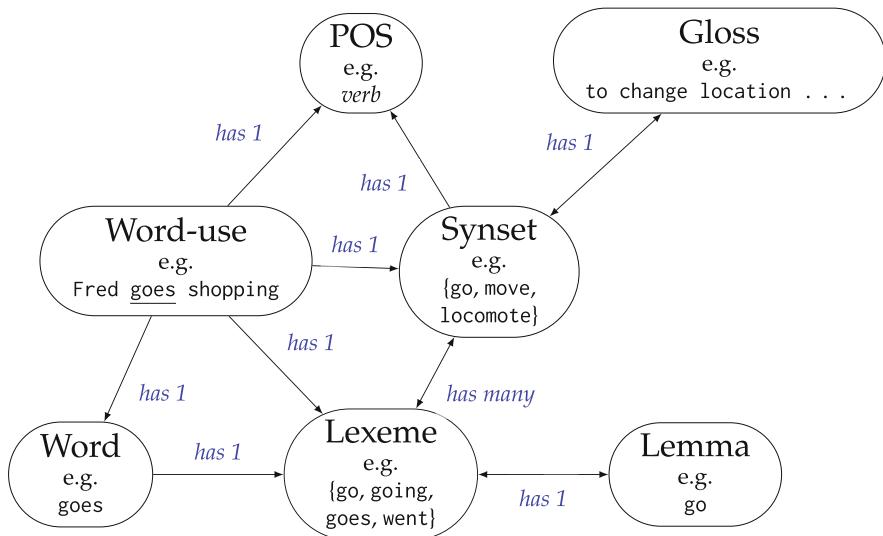
### Unlemmatization

Given lemma (as one can extract from WordNet) and a full POS tag (such as a Brown-style tag) for a word, it is possible to undo the lemmatization with a high degree of reliability using relatively simple rules (again due to the similarity of the lemma to the stem). The POS tag encodes the key inflectional features that are lost. Patten.en (De Smedt and Walter 2012) is a python library encoding such rules (pluralisation, verb conjugation, etc.); though combining them with the POS tag to drive them is a task left for the reader. This can be used to find substitute words using WordNet's features, for finding synonyms, antonyms and other lemmas from lexically related categories.

14 The standard way to assign word senses is via some lexicographical resource, such  
15 as a dictionary, or a thesaurus. There is not a canonical list of word senses that are  
16 consistently defined in English. Every dictionary is unique, with different definitions  
17 and numbers of word senses. The most commonly used lexicographical resource is  
18 WordNet (Miller et al. 1995), and the multi-lingual BabelNet (Navigli and Simone  
19 2010). The relationship between the terminology used in word sense problems is  
shown in Fig. 4.1.

#### 21 **4.1.1 Word Sense Disambiguation**

22 Word sense disambiguation is one of the hardest problems in NLP. Very few systems  
23 significantly outperform the baseline, i.e. the most frequent sense (MFS) technique.  
24 Progress on the problem is made difficult by several factors.



**Fig. 4.1** The relationship between terms used to discuss various word sense problems. The lemma is used as the representation for the lexeme, for WordNet's purposes when indexing. For many tasks each the word-use is pre-tagged with its lemma and POS tag, as these can be found with high reliability using standard tools. Note that the arrows in this diagram are *directional*. That is to say, for example, each Synset *has 1* POS, but each POS *has many* Synsets

LW: move boxes to here \*\*\*

25 The sense is hard to identify from the context. Determining the sense may require  
 26 very long range information: for example the information on context may not even  
 27 be in the same sentence. It may require knowing the domain of the text, because word  
 28 sense uses vary between domains. Such information is external to the text itself. It  
 29 may in-fact be intentionally unclear, with multiple correct interpretations, as in a  
 30 pun. It maybe unintentionally unknowable, due to a poor writing style, such that it  
 31 would confuse any human reader. These difficulties are compounded by the limited  
 32 amount of data available.

33 There is only a relatively small amount of labelled data for word sense prob-  
 34 lems. It is the general virtue of machine learning that given enough data, almost any  
 35 input-output mapping problem (i.e. function approximation) can be solved. Such an  
 36 amount of word sense annotated data is not available. This is in contrast to finding  
 37 unsupervised word embeddings, which can be trained on any text that has ever been  
 38 written. The lack of very large scale training corpora renders fully supervised meth-  
 39 ods difficult. It also results in small sized testing corpora; which leads to systems  
 40 that may appear to perform well (on those small test corpora), but do not generalise  
 41 to real world uses. In addition, the lack of human agreement on the correct sense,  
 42 resulting in weak ground truth, further makes creating new resources harder. This  
 43 limited amount of data compounds the problem's inherent difficulties.

44 It can also be said that word senses are highly artificial and do not adequately repre-  
 45 sent meaning. However, WSD is required to interface with lexicographical resources,

such as translation dictionaries (e.g. BabelNet), ontologies (e.g. OpenCyc), and other datasets (e.g. ImageNet Deng et al. 2009).

It may be interesting to note, that the number of meanings that a word has is approximately inversely proportional related to its frequency of use rank (Zipf 1945). That is to say the most common words have far more meanings than rarer words. It is related to (and compounds with) the more well-known Zipf's Law on word use (Zipf 1949), and can similarly be explained-based on Zipf's core premise of the principle of least effort. This aligns well with our notion that precise (e.g. technical) words exist but are used only infrequently – since they are only explaining a single situation. This also means that by most word-uses are potentially very ambiguous.

The most commonly used word sense (for a given word) is also overwhelmingly more frequent than its less common brethren – word sense usage also being roughly Zipfian distributed (Kilgarriff 2004). For this reason the Most Frequent Sense (MFS) is a surprisingly hard baseline to beat in any WSD task.

LW: Move Lemmatization and Unlemmatization boxes here

### Semantic Syllepsis

(Also known as pathological sentences that kill almost all WSD systems.) Consider the sentence: John used to work for the *newspaper* that you are carrying.. In this sentence the word-use newspaper simultaneously have two different meanings: it is both the company, and the object. This violates our earlier statement that every word-use belongs to exactly one synset. WSD systems are unable to handle these sentences as they attempt to assign a single sense to each word-use. Most word sense induction systems cannot do much better: at best a new sense could be allocated for the joint use, which does not correspond to the linguistic notion of the word having two senses for different parts of the sentence. Most works on word sense disambiguation outright ignore these sentences, or consider them to be ungrammatical, or incorrect. However, they are readily understood and used without thought by most native speakers. These constructions are also known as *zeugma*, although zeugma is itself a highly polysemous word, so its usage varies.

#### 4.1.1.1 Most Frequent Sense

Given a sense annotated corpus, it is easy to count how often each sense of a word occurs. Due to the over-whelming frequency of the most frequent sense, it is unlikely for even a small training corpus to have the most frequent sense differing from the use in the language as a whole.

The Most Frequent Sense (MFS) method of word sense disambiguation is defined by counting the frequency of a particular word sense for a particular POS tagged word. For the  $i$ th word use being the word  $w^i$ , having some sense  $s^j$  then without any further context the probability of that sense being the correct sense is  $P(s^j | w^i)$ . One can use the part of speech tag  $p_i$  (for the  $i$ th word use) as an additional condition, and thus find  $P(s^j | w^i, p_i)$ . WordNet encodes this information for each lemma-synset

pair (i.e. each word sense) using the SemCor corpus counts. This is also used for sense ordering, which is why most frequent sense is sometimes called first sense. This is a readily available and practical method for getting a baseline probability of each sense. Most frequent sense can be applied for word sense disambiguation using this frequency-based probability estimate:  $\text{argmax}_{\forall s,j} P(s^j \mid w^i, p_i)$ .

In the most recent SemEval WSD task (Moro and Roberto 2015), MFS beat all submitted entries for English, both overall, and on almost all cuts of the data. The results for other languages were not as good, however in other languages the true corpus-derived sense counts were not used.

### WordNet is not a strong moral baseline

WordNet, as a resource, based partly on the work of Princeton undergraduate students in the early 1990s, and on the literature of 1961, is not the kind of resource one might hope for from an AI information perspective. The glosses include a number of biases. These biases are reflective of the language use, but are not necessarily ideal to be encoded into a system. For example: S: (v) nag, peck, hen-peck (bother persistently with trivial complaints) ‘She nags her husband all day long’s. Other dictionaries regularly show up in the News for similar content.

Another problem is the source of the word sense counts. As discussed in the main text, sense counts are important in WSD systems. The counts come from SemCor, a sense annotated subset of the Brown Corpus. The Brown Corpus is a sampling of American texts from 1961. The cultural norms of 1961 were not the norms of today. (For context, note that the US did not pass the Civil Rights act to end segregation until 1964). As such, one should not trust WordNet (or SemCor) to reflect current sense counts, for words which have undergone usage change since 1961.

Furthermore, when creating down stream resources-based on WordNet, one should not use these sense counts to determine how important it is to include a concept. If ImageNet (Deng et al. 2009) for example, had used SemCor counts to determine which synsets of images would be included, then items rarely discussed in 1961 literature, like wheelchairs, and prosthesis would be excluded. Which would in turn make many image processing systems systematically unhelpful in processing images relating to the disabled. (Do not fear: even the initial release of ImageNet contains hundreds of images of wheelchairs, and prosthesis) Unintentional biasing of data can have on-going effects on the behaviour of machine learning-based systems far beyond the original conception.

## 4.2 Word Sense Representation

It is desirable to create a vector representation of a word sense much like in this chapter representations were created for words. We desire to an embedding to represent each word sense, as normally represented by a word-synset pair. This section considers the representations for the lexical word senses as given from a dictionary. We consider a direct method of using a labelled corpus, and an indirect method makes use of simpler sense-embeddings to partially label a corpus before retraining. These methods create representations corresponding to senses from WordNet. Section 4.3 considers the case when the senses are to also be discovered, as well as represented.

### 4.2.1 Directly Supervised Method

The simple and direct method is to take a dataset that is annotated with word senses, and then treat each sense-word pair as if it were a single word, then apply any of the methods for word representation discussed in this chapter. Iacobacci et al. (2015) use a CBOW language model (Mikolov et al. 2013) to do this. This does, however, run into the aforementioned problem, that there is relatively little training data that has been manually sense annotated. Iacobacci et al. (2015) use a third-party WSD tool, namely BabelFly (Moro et al. 2014), to annotate the corpus with senses. This allows for existing word representation techniques to be applied.

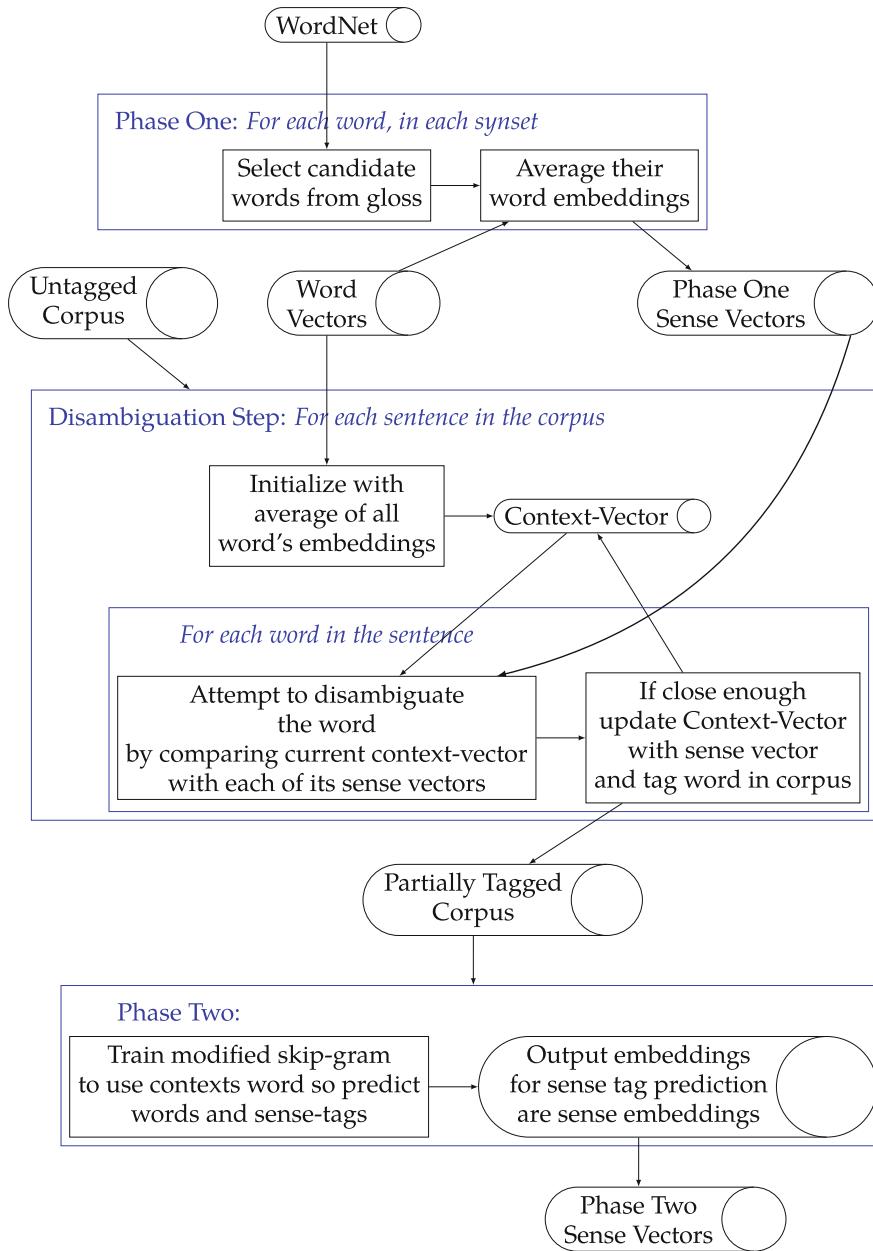
Chen et al. (2014) applies a similar technique, but using a word-embedding-based partial WSD system of their own devising, rather than an external WSD tool.

### 4.2.2 Word Embedding-Based Disambiguation Method

Chen et al. (2014) uses an almost semi-supervised approach to train sense vectors. They partially disambiguate their training corpus, using initial word sense vectors and WordNet. They then completely replace these original (phase one) sense-vectors, by using the partially disambiguated corpus to train new (phase two) sense-vectors via a skip-gram variant. This process is shown in Fig. 4.2.

The **first phase** of this method is in essence a word-embedding-based WSD system. When assessed as such, they report that it only marginally exceeds the MFS baseline, though that is not at all unusual for WSD algorithms as discussed above.

They assign a sense vector to every word sense in WordNet. This sense vector is the average of word-embeddings of a subset of words in the gloss, as determined using pretrained skip-grams (Mikolov et al. 2013). For the word  $w$  with word sense  $w^{s^i}$ , a set of candidate words,  $cands(w^{s^i})$ , is selected from the gloss based on the following set of requirements. First, the word must be a content word: that is a verb, noun, adverb or adjective; secondly, its cosine distance to  $w$  must be below some threshold  $\delta$ ; finally, it must not be the word itself. When these requirements are followed  $cands(w^{s^i})$  is a set of significant closely related words from the gloss.



**Fig. 4.2** The process used by Chen et al. (2014) to create word sense embeddings.

### WSD with embeddings

It is beyond the scope of this work to fully discuss WSD systems. However, we will remark that (single sense) word embeddings are a generally useful feature as an input to any NLP ML system. As such they can be used as features in a fully supervised WSD system. The idea of using them in this way is similar to the LSI enhanced Lesk WSD system of Basile et al. (2014).

### Cosine distance

Here we talk of cosine distance, where a smaller distance implies more similar (and 0.0 identical). Contrasting this with the cosine similarity, where higher value implies more similar (and 1.0 identical).

Cosine distance is still not a true metric as  $d^{\cos}(v, kv) = 0$  for all  $k \in \mathbb{R}_+$ . Other times you may see cosine similarity, ranging between  $-1$  (most different) and  $1$  (most similar). Cosine similarity is given by  $\text{sim}(a, b) = \frac{\tilde{a} \cdot \tilde{b}}{\|\tilde{a}\|_2 \|\tilde{b}\|_2} =$

*LW: make this equation not split between lines*  
 $\cos(\angle \tilde{a} \tilde{b})$  i.e. the unit-length normalised dot product of the vectors. Cosine distance is usually defined as  $d^{\cos}(\tilde{a}, \tilde{b}) = \frac{1 - \text{sim}(\tilde{a}, \tilde{b})}{2}$ . Ranging between 0 (most similar) and 1 (most different).

The phase one sense vector for  $w^{s^i}$  is the mean of the word vectors for all the words in  $cands(w^{s^i})$ . The effect of this is that we expect that the phase one sense vectors for most words in the same synset will be similar but not necessarily identical. This expectation is not guaranteed however. As an example, consider the use of the word `china` as a synonym for `porcelain`: the single sense vector for `china` will likely be dominated by its more significant use referring to the country, which would cause very few words in the gloss for the `porcelain` synset to be included in  $cands$ . Resulting in the phase one sense vectors for the synonymous senses of `porcelain` and `china` actually being very different.

The phase one sense vectors are used to disambiguate the words in their unlabelled training corpus. For each sentence in the corpus, an initial *content vector* is defined by taking the mean of the skip-gram word embedding (not word sense) for all content words in the sentence. For each word in the sentence, each possible sense-embedding is compared to the context vector. If one or more senses vectors are found to be closer than a given threshold, then that word is tagged with the closest of those senses, and the context vector is updated to use the sense-vector instead of the word vector. Words that do not come within the threshold are not tagged, and the context vector is not updated. This is an important part of their algorithm, as it ensures that words without clear senses do not get a sense ascribed to them. This thus avoids any dubious sense tags for the next training step.

In **phase two** of training (Chen et al. 2014) employ the skip-gram word-embedding method, with a variation, to predict the word senses. They train it on the partially disambiguated corpus produced in phase one. The original sense vectors are dis-

140 carded. Rather than the model being tasked only to predict the surrounding words,  
141 it is tasked to predict surrounding words and their sense-tags (where present). In the  
142 loss function the prediction of tags and words is weighted equally.

143 Note that the input of the skip-gram is the just central word, not the pair of central  
144 word with sense-tag. In this method, the word sense embeddings are output embed-  
145 dings; though it would not be unreasonable to reverse it to use input embeddings  
146 with sense tags, or even to do both. The option to have input embeddings and out-  
147 put embeddings be from different sets, is reminiscent of Schwenk (2004) for word  
148 embeddings.

149 The phase one sense vectors have not been assessed on their representational  
150 quality. It could be assumed that because the results for these were not reported, they  
151 were worse than those found in phase two. The phase two sense vectors were not  
152 assessed for their capacity to be used for word sense disambiguation. It would be  
153 desirable to extend the method of Chen et al. (2014), to use the phase two vectors  
154 for WSD. This would allow this method to be used to disambiguate its own training  
155 data, thus enabling the method to become self-supervised.

### 156 4.3 Word Sense Induction (WSI)

#### Can we go from induced senses to lexical senses?

A natural question given the existence of many WSI systems, and the existing wealth of lexically indexed resources, is if we can align induced senses to a set of lexically defined senses. Agirre et al. (2006) proposed a method for doing this using a weighted mapping-based on the probabilities found using induced sense WSD on a labelled “mapping” corpus. This has only been used on relatively small datasets with only hundreds of words (SenseEval 3 Mihalcea et al. 2004 and SemEval-2007 Task 02 Agirre and Aitor 2007). Our own investigations in White et al. (2018) with the larger SemEval 2007 Task 7 (Navigli et al. 2007) suggest that it may not scale very well to real-word WSD tasks. That work proposed an alternative method that worked better, though still not as well as could be hoped. Finding suitable methods to link unsupervised representations, to human defined senses remains a topic worthy of research.

157 In this section we will discuss methods for finding a word sense without reference  
158 to a standard set of senses. Such systems must discover the word senses at the same  
159 time as they find their representations. One strong advantage of these methods is  
160 that they do not require a labelled dataset. As discussed there are relatively few high-  
161 quality word sense labelled datasets. The other key advantage of these systems is that  
162 they do not rely on fixed senses determined by a lexicographer. This is particularly  
163 useful if the word senses are highly domain specific; or in a language without strong  
164 lexicographical resources. This allows the data to inform on what word senses exist.

### Why do skip-grams perform so well on SCWS?

SCWS is a corpus designed for evaluating word sense embeddings. Single sense embeddings (e.g. skip-grams) cannot take advantage of the context information in the SCWS. However, they do often perform comparably to the word sense embeddings. Sometimes even outperforming them. It is unclear if this highlights the difficulty of the task (i.e. that the impact of context is hard to gauge), or it might be due to the (implicit) most frequent sense dominating both the use in the tasks, and the representation in a single sense. Alternatively, it may just be the result of the fine tuning of the more mature single-sense embedding methods (and that with more time and tuning multiple sense methods could do proportionally better).

165 Most vector word sense induction and representation approaches are evaluated  
166 on similarity tests. Such tests include WordSim-353 (Finkelstein et al. 2001) for  
167 context-less, or Stanford’s Contextual Word Similarities (SCWS) for similarity with  
168 context information (Huang et al. 2012). This evaluation is also suitable for evaluating  
169 single sense word-embeddings, e.g. skip-grams.

170 We can divide the WSI systems into context clustering-based approaches, and co-  
171 location prediction-based approaches. This division is similar to the separation of co-  
172 location matrix factorisation, and co-location prediction-based approaches discussed  
173 in this chapter. It can be assumed thus that at the core, like for word embeddings, they  
174 are fundamentally very similar. One could think of prediction of collocated words  
175 as a soft indirect clustering of contexts that can have those words.

#### 176 4.3.1 Context Clustering-Based Approaches

177 As the meaning of a word, according to word embedding principles, is determined  
178 by the contexts in which it occurs, we expect that different meanings (senses) of the  
179 same words should occur in different contexts. If we cluster the contexts that a word  
180 occurs in, one would expect to find distinct clusters for each sense of the word. It is  
181 on this principle that the context clustering-based approaches function.

##### 182 4.3.1.1 Offline Clustering

183 The fundamental method for most clustering-based approaches is as per Schütze  
184 (1998). That original work is not a neural word sense embedding, however the  
185 approach remains the same. Pantel and Dekang (2002) and Reisinger and Raymond  
186 (2010) are also not strictly neural word embedding approaches (being more classical  
187 vector representations), however the overall method is also very similar.

#### On context representations

LW: Move this box such that it is not split across pages

These co-location clustering methods require finding a representation for the  
context (from this a similarity metric is applied, and the clustering is then

done). More generally, this can be related to the next chapter: Chap. 6, as any of these methods could be used to derive a vector representation of a context. In most works (including all the works discussed here) comparatively simple representations of the contexts are used. It would be interesting to extend the sentence representation methods, and apply them to this use.

The clustering process is done by considering all word uses, with their contexts. The contexts can be a fixed-sized window of words (as is done with many word-embedding models), the sentence, or defined using some other rule. Given a pair of contexts, some method of measuring their similarity must be defined. In vector representational works, this is ubiquitously done by assigning each context a vector, and then using the cosine similarity between those vectors.

The first step in all the offline clustering methods is thus to define the representations of the contexts. Different methods define the context vectors differently:

- Schütze (1998) uses variations of inverse-document-frequency (idf) weighted bags of words, including applying dimensionality reduction to find a dense representation.
- Pantel and Dekang (2002) use the mutual information vectors between words and their contexts.
- Reisinger and Raymond (2010), use td-idf or  $\chi^2$  weighted bag of words.
- Huang et al. (2012) uses td-idf weighted averages of (their own) single sense word embeddings for all words in the context.
- Kågebäck et al. (2015) also uses a weighted average of single sense word skip-gram embeddings, with the weighting based on two factors. One based on how close the words were, and the other on how likely the co-occurrence was according to the skip-gram model.

It is interesting to note that idf, td-idf, mutual information, skip-gram co-occurrence probabilities (being a proxy for point-wise mutual information Levy and Yoav 2014), are all closely related measures.

**LW: Try to move this box so it is not split across pages, if possible**  
**On clustering**

Clustering can be defined as a (mixed integer) optimisation task, of assigning points to clusters so as to satisfy some loss function-based around minimising intra-cluster variance while maximising inter-cluster variance (or a similar measure). As this is NP-hard, most clustering methods are approximate. K-means is very popular because of its simplicity, however it easily falls into local minima, and so normally it is run dozens of times (at least) to obtain more optimal results. K-means also has the issue of having to select the number of clusters ( $k$ ). It should be remembered that there exist many other clustering methods than k-means (and its variants). These other methods use different loss functions, and different strategies to overcome the NP-hard nature of the problem. In particular their mixture model methods, hierarchical methods, spectral methods, and others. We personally favour affinity propa-

gation (Frey and Delbert 2007), though there is provably no ideal clustering algorithm even in the non-heuristic case (Kleinberg 2003). On any clustering task (word sense or otherwise) it is worth investigating several clustering algorithms, and not just settling for k-means (particularly not settling for k-means run once.). A series of interesting and easy reading articles on clustering can be found at: <http://alexhwilliams.info/itsneuronalblog/2015/09/11/clustering1/>, <http://alexhwilliams.info/itsneuronalblog/2015/10/01/clustering2/>, and <http://alexhwilliams.info/itsneuronalblog/2015/11/18/clustering-is-easy/>

211 The **second step** in off-line clustering is to apply a clustering method to cluster the  
212 word-uses. This clustering is done based on the calculated similarity of the context  
213 representation where the words are used. Again, different WSI methods use different  
214 clustering algorithms.

- 215 • Schütze (1998) uses a group average agglomerative clustering method.  
216 • Pantel and Dekang (2002) use a custom hierarchical clustering method.  
217 • Reisinger and Raymond (2010) use mixtures of von-Mises-Fisher distributions.  
218 • Huang et al. (2012) use spherical k-means.  
219 • Kågebäck et al. (2015) use k-means.

220 The **final step** is to find a vector representation of each cluster. For non-neural  
221 embedding methods this step is not always done, as defining a representation is not  
222 the goal, though in general it can be derived from most clustering techniques. Schütze  
223 (1998) and Kågebäck et al. (2015) use the centroids of their clusters. Huang et al.  
224 (2012) use a method of relabelling the word uses with a cluster identifier, then train  
225 a (single-sense) word embedding method on cluster identifiers rather than words.  
226 This relabelling technique is similar to the method later used by Chen et al. (2014)  
227 for learning lexical sense representations, as discussed in Sect. 4.2.2. As each cluster  
228 of contexts represents a sense, those cluster embeddings are thus also considered as  
229 suitable word sense embeddings.

230 To summarize, all the methods for inducing word sense embeddings via off-line  
231 clustering follow the same process. **First:** represent the contexts of word use, so as  
232 to be able to measure their similarity. **Second:** use the context's similarity to cluster  
233 them. **Finally:** find a vector representation of each cluster. This cluster representation  
234 is the induced sense embedding.

### 235 4.3.1.2 Online Clustering

236 The methods discussed above all use off-line clustering. That is to say the clustering  
237 is performed after the embedding is trained. Neelakantan et al. (2015) perform the  
238 clustering during training. To do this they use a modified skip-gram-based method.  
239 They start with a fixed number of randomly initialised sense vectors for each con-  
240 text. These sense vectors are used as input embeddings for the skip-gram context  
241 prediction task, over single sense output embeddings. Each sense also has, linked to

it, a context cluster centroid, which is the average of all output embeddings for the contexts that the sense is assigned to. Each time a training instance is presented, the average of the context output embeddings is compared to each sense's context cluster centroid. The context is assigned to the cluster with the closest centroid, updating the centroid value. This can be seen as similar to performing a single k-means update step for each training instance. Optionally, if the closest centroid is further from the context vector than some threshold, a new sense can be created using that context vector as the initial centroid. After the assignment of the context to a cluster, the corresponding sense vector is selected for use as the input vector in the skip-gram context prediction task.

Kågebäck et al. (2015) investigated using their weighting function (as discussed in Sect. 5.3.1.1) with the online clustering used by Neelakantan et al. (2015). They found that this improved the quality of the representations. More generally any such weighting function could be used. This online clustering approach is loosely similar to the co-location prediction-based approaches.

### 4.3.2 Co-location Prediction-Based Approaches

#### Probability

One may wish to brush up on basic probability notions for this section. In particular joint, conditional and marginal probabilities definitions; as well as Bayes Theorem and the probability chain-rule which come from those. In brief these are as follows.

**Conditional Probability:** LW, unbold remove line break. Merge with box above  
 $P(A | B) = \frac{P(A, B)}{P(B)}$

**Marginal Probability:** LW, unbold remove line break. Merge with box above  
 $P(A) = \sum_{\forall b} P(A, B = b)$

**Bayes Theorem:** LW, unbold remove line break. Merge with box above  
 $P(A | B) = \frac{P(B|A)P(A)}{P(B)}$

**Probability Chain-rule:** LW, unbold remove line break. Merge with box above  
 $P(A^n, \dots, A^1) = P(A^n | A^{n-1}, \dots, A^1)P(A^{n-1}, \dots, A^1)$   
e.g.

$$\begin{aligned} P(A, B, C) \\ = P(A \mid B, C)P(B \mid C)P(C) \end{aligned}$$

### LW:Merge with box above

The latter three rules are consequences of the first.

Rather than clustering the contexts, and using those clusters to determine embeddings for different senses, one could consider the sense as a latent variable in the task used to find word embeddings – normally a language modelling task. The principle is that it is not the word that determines its collocated context words, but rather the word sense. So the word sense can be modelled as a hidden variable, where the word, and the context words are being observed.

Tian et al. (2014) used this to define a skip-gram-based method for word sense embeddings. For input word  $w^i$  with senses  $\mathcal{S}(w^i)$ , the probability of output word  $w^o$  occurring near  $w^i$  can be given as:

$$P(w^o \mid w^i) = \sum_{\forall s^k \in \mathcal{S}(w^i)} P(w^o \mid s^k, w^i)P(s^k \mid w^i) \quad (4.1)$$

Given that a sense  $s^k$  only belongs to one word  $w^i$ , we know that  $k$ th sense of the  $i$ th word only occurs when the  $i$ th word occurs. We have that the joint probability  $P(w^i, s^k) = P(s^k)$ .

We can thus rewrite Eq. 4.1 as:

$$P(w^o \mid w^i) = \sum_{\forall s^k \in \mathcal{S}(w^i)} P(w^o \mid s^k)P(s^k \mid w^i) \quad (4.2)$$

A softmax classifier can be used to define  $P(w^o \mid s^k)$ , just like in normal language modelling. With output embeddings for the words  $w^o$ , and input embeddings for the word senses  $s^k$ . This softmax can be sped-up using negative sampling or hierarchical softmax. The later was done by Tian et al. (2014).

Equation 4.2 is in the form of a mixture model with a latent variable. Such a class of problems are often solved using the Expectation Maximisation (EM) method. In short, the EM procedure functions by performing two alternating steps. The **E-step** calculates the expected chance of assigning word sense for each training case ( $\hat{P}(s^l \mid w^o)$ ) in the training set  $\mathcal{X}$ . Where a training case is a pairing of a word use  $w^i$ , and context word  $w^o$ , with  $s^l \in \mathcal{S}(w^i)$ , formally we have:

$$\hat{P}(s^l \mid w^o) = \frac{\hat{P}(s^l \mid w^i)P(w^o \mid s^l)}{\sum_{\forall s^k \in \mathcal{S}(w^i)} \hat{P}(w^o \mid s^k)P(s^k \mid w^i)} \quad (4.3)$$

The **M-step** updates the prior likelihood of each sense (that is without context) using the expected assignments from the E-step.

$$\hat{P}(s^l \mid w^i) = \frac{1}{|\mathcal{X}|} \sum_{\forall (w^o, w^i) \in \mathcal{X}} \hat{P}(s^l \mid w^o) \quad (4.4)$$

During this step the likelihood of the  $P(w^o | w^i)$  can be optimised to maximise the likelihood of the observations. This is done via gradient descent on the neural network parameters of the softmax component:  $P(w^o | s^k)$ . By using this EM optimisation the network can fit values for the embeddings in that softmax component.

A limitation of the method used by Tian et al. (2014), is that the number of each sense must be known in advance. One could attempt to solve this by using, for example, the number of senses assigned by a lexicographical resource (e.g. WordNet). However, situations where such resources are not available or not suitable are one of the main circumstances in which WSI is desirable (for example in work using domain specific terminology, or under-resourced languages). In these cases one could apply a heuristic-based on the distribution of senses-based on the distribution of words (Zipf 1945). An attractive alternative would be to allow senses to be determined-based on how the words are used. If they are used in two different ways, then they should have two different senses. How a word is being used can be determined by the contexts in which it appears.

Bartunov et al. (2015) extend on this work by making the number of senses for each word itself a fit-able parameter of the model. This is a rather Bayesian modelling approach, where one considers the distribution of the prior.

### WordNet and BabelNet

As mentioned in the previous sections, WordNet and BabelNet are the predominant lexicons used for word senses. It is not directly relevant to this section, but we have space here to remark upon them. WordNet (Tengi et al. 1998) as a very well established tool has a binding in practically every modern programming language suitable for NLP. WordNet.jl (<https://github.com/JuliaText/WordNet.jl>) is the Julia binding. NLTK (Bird et al. 2009) includes one for Python. BabelNet (Navigli and Simone 2010) is intended to be accessed as an online resource, via a RESTful API. Users receive 1000 free queries per day. Academic users can request an upgrade to 50,000 queries per day, or to download a copy of the database. From personal experience we found those requests to be handled easily and rapidly.

Considering again the form of Eq. 4.2

$$P(w^o | w^i) = \sum_{\forall s^k \in \mathcal{S}(w^i)} P(w^o | s^k) P(s^k | w^i) \quad (4.5)$$

The prior probability of a sense given a word, but no context, is  $P(s^k | w^i)$ . This is Dirichlet distributed. This comes from the definition of the Dirichlet distribution as the prior probability of any categorical classification task. When considering that the sense may be one from an unlimited collection of possible senses, then that prior becomes a Dirichlet process.

In essence, this prior over a potentially unlimited number of possible senses becomes another parameter of the model (along with the input sense embeddings and output word embeddings). The fitting of the parameters of such a model is beyond

the scope of this book; it is not entirely dissimilar to the fitting via expectation maximisation incorporating gradient descent used by Tian et al. (2014). The final output of Bartunov et al. (2015) is as desired: a set of induced sense embeddings, and a language model that is able to predict how likely a word is to occur near that word sense ( $P(w^o | s^k)$ ).

By application of Bayes' theorem, the sense language model can be inverted to take a word's context, and predict the probability of each word sense.

### Independence Assumption

Technically, Eq. 4.6 does not require the independence of the probabilities of the context words. Rather it only requires that the context words be conditionally independent on the word in question  $w^i$ . Nevertheless, even the conditional independence assumption is incorrect, except for a theoretical perfect embedding capturing perfect information. The conditional independence assumption remains useful as an approximation.

$$P(s^l | w^o) = \frac{P(w^o | s^l)P(s^l | w^i)}{\sum_{\forall s^k \in \mathcal{S}(w^i)} P(w^o | s^k)P(s^k | w^i)} \quad (4.6)$$

with the common (but technically incorrect) assumption that all words in the context are independent.

LW: Move to after conclusion. At end of chapter

### Finding the nearest neighbours (Nearest Neighbour Trees)

A common evaluation task with any representation is to find its nearest neighbours. The naïve solution is to check the distance to all points. For  $n$  points this is  $O(n)$  operations. For word embeddings  $n$  is the size of the vocabulary, perhaps 100,000 words. Performing 100,000 operations per check, is not entirely unreasonable on modern computers (even when the operations are on 300 dimensional representations). However, for word sense embeddings, which have many senses per word in the vocabulary, this means many more points to check. 30 senses per word is not unusual for fine-grained word sense induction. Having a total  $n = 3,000,000$  representations to check causes a noticeable delay. To solve this we can use data structures designed for fast nearest neighbour querying. A k-d tree takes at worst  $O(n \log_2(n))$  time to construct. Once constructed on average it takes  $O(\log(n))$  to find the nearest neighbour to any point. This makes checking the nearest neighbour nearly instantaneous for even the largest vocabularies.

Given a context window:  $\mathcal{W}^i = (w^{i-\frac{n}{2}}, \dots, w^{i-1}, w^{i+1}, \dots, w^{i+\frac{n}{2}})$ , we have:

$$P(s^l | \mathcal{W}^i) = \frac{\prod_{\forall w^j \in \mathcal{W}^i} P(w^j | s^l)P(s^l | w^i)}{\sum_{s^k \in \mathcal{S}(w^i)} \prod_{\forall w^j \in \mathcal{W}^i} P(w^j | s^k)P(s^k | w^i)} \quad (4.7)$$



## 331 4.4 Conclusion

332 Word sense representations allow the representations of the senses of words when one  
 333 word has multiple meanings. This increases the expressiveness of the representation.  
 334 These representations can in general be applied anywhere word embeddings can.  
 335 They are particularly useful for translation, and in languages with large numbers of  
 336 homonyms.

337 The word representation discussions in this chapter naturally lead to the next  
 338 section on phrase representation. Rather than a single word having many meanings,  
 339 the next chapter will discuss how a single meaning may take multiple words to  
 340 express. In such longer structure's representations, the sense embeddings discussed  
 341 here are often unnecessary, as the ambiguity may be resolved by the longer structure.  
 342 Indeed, the methods discussed in this chapter have relied on that fact to distinguish  
 343 the senses using the contexts.

344 LW: Move nearest neighbours box here.

## 345 References

- 346 Agirre, Eneko and Aitor Soroa. 2007. Semeval-2007 task 02: Evaluating word sense induction and  
 347 discrimination systems. In *Proceedings of the 4th international workshop on semantic evalua-*  
 348 *tions. SemEval '07*, 7–12. Prague, Czech Republic: Association for Computational Linguistics.
- 349 Agirre, Eneko, David Martínez, Oier López De Lacalle, and Aitor Soroa. 2006. Evaluating and  
 350 optimizing the parameters of an unsupervised graph-based WSD algorithm. In *Proceedings of*  
 351 *the first workshop on graph based methods for natural language processing*, 89–96. Association  
 352 for Computational Linguistics.
- 353 Bartunov, Sergey, Dmitry Kondrashkin, Anton Osokin, and Dmitry P. Vetrov. 2015. Breaking sticks  
 354 and ambiguities with adaptive skip-gram. In *CoRR*. [arXiv:1502.07257](https://arxiv.org/abs/1502.07257).
- 355 Basile, Pierpaolo, Annalina Caputo, and Giovanni Semeraro. 2014. An Enhanced lesh word sense  
 356 disambiguation algorithm through a distributional semantic model. In *Proceedings of COLING*  
 357 *2014, the 25th international conference on computational linguistics: Technical papers. Dublin*,  
 358 1591–1600. Ireland: Dublin City University and Association for Computational Linguistics.
- 359 Bird, Steven, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python*.  
 360 O'Reilly Media, Inc.
- 361 Chen, Xinxiong, Zhiyuan Liu, and Maosong Sun. 2014. A unified model for word sense represen-  
 362 tation and disambiguation. In *EMNLP* (Citeseer), 1025–1035.
- 363 De Smedt, Tom and Walter Daelemans. 2012. Pattern for python. *The Journal of Machine Learning*  
 364 *Research* 13 (1): 2063–2067.
- 365 Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A large-scale  
 366 hierarchical image database. In *CVPR09*.
- 367 Finkelstein, Lev, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and  
 368 Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th*  
 369 *international conference on World Wide Web*, 406–414. ACM.
- 370 Frey, Brendan J., and Delbert Dueck. 2007. Clustering by passing messages between data points.  
 371 *Science* 315 (5814): 972–976.
- 372 Huang, Eric H., Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2012. Improving  
 373 word representations via global context and multiple word proto-types. In *Proceedings of the 50th*  
 374 *annual meeting of the association for computational linguistics: Long papers*, vol. 1, 873–882.  
 375 Association for Computational Linguistics.

- Iacobacci, Ignacio, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. SensEmbed: Learning sense embeddings for word and relational similarity. In *Proceedings of ACL*, 95–105.
- Kågebäck, Mikael, Fredrik Johansson, Richard Johansson, and Devdatt Dubhashi. 2015. Neural context embeddings for automatic discovery of word senses. In *Proceedings of NAACL-HLT*, 25–32.
- Kilgarriff, Adam. 2004. How dominant is the commonest sense of a word? In *Text, speech and dialogue: 7th international conference, TSD 2004, Brno, Czech Republic, September 8–11, 2004. Proceedings*, eds. Petr Sojka, Ivan Kopecek, and Karel Pala, 103–111. Berlin, Heidelberg: Springer. ISBN: 978-3-540-30120-2. [https://doi.org/10.1007/978-3-540-30120-2\\_14](https://doi.org/10.1007/978-3-540-30120-2_14).
- Kleinberg, Jon M. 2003. An impossibility theorem for clustering. In *Advances in neural information processing systems*, 463–470.
- Levy, Omer and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems*, 2177–2185.
- Mihalcea, Rada, Timothy Anatolievich Chklovski, and Adam Kilgarriff. 2004. The senseval-3 english lexical sample task. In *Association for computational linguistics*.
- Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. [arXiv:1301.3781](https://arxiv.org/abs/1301.3781).
- Miller, George A. 1995. WordNet: A lexical database for English. *Communications of the ACM* 38 (11): 39–41.
- Moro, Andrea and Roberto Navigli. 2015. SemEval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In *Proceedings of SemEval- 2015*.
- Moro, Andrea, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: A unified approach. *Transactions of the Association for Computational Linguistics (TACL)* 2: 231–244.
- Navigli, Roberto and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, 216–225. Association for Computational Linguistics.
- Navigli, Roberto, Kenneth C. Litkowski, and Orin Hargraves. 2007. SemEval- 2007 task 07: Coarse-grained english all-words task. In *Proceedings of the 4th international workshop on semantic evaluations. SemEval '07*, 30–35. Prague, Czech Republic: Association for Computational Linguistics.
- Neelakantan, Arvind, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2015. Efficient non-parametric estimation of multiple embeddings per word in vector space. [arXiv:1504.06654](https://arxiv.org/abs/1504.06654).
- Pantel, Patrick and Dekang Lin. 2002. Discovering word senses from text. In *Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining*, 613–619. ACM.
- Reisinger, Joseph and Raymond J. Mooney. 2010. Multi-prototype vector-space models of word meaning. In *Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics*, 109–117. Association for Computational Linguistics.
- Schütze, Hinrich. 1998. Automatic word sense discrimination. *Computational Linguistics*. 24 (1): 97–123. ISSN: 0891-2017.
- Schwenk, Holger. 2004. Efficient training of large neural networks for language modeling. In *2004 IEEE international joint conference on neural networks, 2004. Proceedings*, 4, 3059–3064. IEEE.
- Tengi, Randee I. 1998. *Design and implementation of the WordNet lexical database and searching software. WordNet: An electronic lexical database*, ed. Christiane (réd.) Fellbaum, 105. Cambridge: The MIT Press.
- Tian, Fei, Hanjun Dai, Jiang Bian, Bin Gao, Rui Zhang, Enhong Chen, and Tie- Yan Liu. 2014. A probabilistic model for learning multi-prototype word embeddings. In *COLING*, 151–160.

- 425 White, Lyndon, Roberto Tognoni, Wei Liu, and Mohammed Bennamoun. 2018. Finding word sense  
426 embeddings of known meaning. In *19th international conference on intelligent text processing*  
427 and computational linguistics (CICLing).
- 428 Zipf, George Kingsley. 1945. The meaning-frequency relationship of words. *The Journal of General*  
429 *Psychology* 33 (2): 251–256.
- 430 Zipf, G.K. 1949. *Human behavior and the principle of least effort: An introduction to human*  
431 *ecology*. Cambridge: Addison-Wesley Press.

# Author Queries

Chapter 4

---

Query Refs.	Details Required	Author's response
	No queries.	

# Metadata of the chapter that will be visualized in SpringerLink

Book Title	Neural Representations of Natural Language	
Series Title		
Chapter Title	Sentence Representations and Beyond	
Copyright Year	2019	
Copyright HolderName	Springer Nature Singapore Pte Ltd.	
Corresponding Author	Family Name Particle Given Name Prefix Suffix Role Division Organization Address Email	<b>White</b> <b>Lyndon</b> <b>Togneri</b> <b>Roberto</b> <b>Liu</b> <b>Wei</b> <b>Bennamoun</b> <b>Mohammed</b>
Author	School of Electrical, Electronic and Computer Engineering The University of Western Australia Perth, WA, Australia LYNDON.WHITE@RESEARCH.UWA.EDU.AU	
Author	School of Electrical, Electronic and Computer Engineering The University of Western Australia Perth, WA, Australia	
Author	School of Computer Science and Software Engineering The University of Western Australia Perth, WA, Australia	
Author		

Role  
Division School of Computer Science and Software Engineering  
Organization The University of Western Australia  
Address Perth, WA, Australia  
Email

---

Abstract	This chapter discusses representations for larger structures in natural language. The primary focus is on the sentence level. However, many of the techniques also apply to sub-sentence structures (phrases), and super-sentence structures (documents). The three main types of representations discussed here are: unordered models, such as sum of word embeddings; sequential models, such as recurrent neural networks; and structured models, such as recursive autoencoders.
----------	--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

---



# Sentence Representations and Beyond

5

*A sentence is a group of words expressing a complete thought*

LW: Extra linespace is added here, that is not in other epigraphs

English Composition and Literature,  
Webster, 1923

- 1 This chapter discusses representations for larger structures in natural language. The  
2 primary focus is on the sentence level. However, many of the techniques also apply  
3 to sub-sentence structures (phrases), and super-sentence structures (documents). The  
4 three main types of representations discussed here are: unordered models, such as  
5 sum of word embeddings; sequential models, such as recurrent neural networks; and  
6 structured models, such as recursive autoencoders.

## Word Embeddings as a by-product

Many sentence representation methods produce word embeddings as a by-product. These word embeddings are either output embeddings, from the softmax, or input embeddings from a lookup layer.

## Initialising input embeddings

It is common (but not ubiquitous) to initialise the input embeddings using pretrained embeddings from one of the methods discussed in Chap. 4, then allow them to be fine-tuned while training the sentence representation method.

- 7 It can be argued that the core of true AI, is in capturing and manipulating the rep-  
8 resentation of an idea. In natural language a sentence (as defined by Webster in the  
9 quote above), is such a representation of an idea, but it is not machine manipulatable.

© Springer Nature Singapore Pte Ltd. 2019

93

L. White et al., *Neural Representations of Natural Language*,  
Studies in Computational Intelligence 783,  
[https://doi.org/10.1007/978-981-13-0062-2\\_5](https://doi.org/10.1007/978-981-13-0062-2_5)



10 As such the conversion of sentences to a machine manipulatable representation is an  
 11 important task in AI research.

12 All techniques which can represent documents (or paragraphs) by necessity rep-  
 13 resent sentences as well. A document (or a paragraph), can consist only of a sin-  
 14 gle sentence. Many of these models always work for sub-sentence structures also,  
 15 like key-phrases. When considering representing larger documents, neural network  
 16 embedding models directly compete with vector information retrieval models, such  
 17 as LSI (Dumais et al. 1988), probabilistic LSI (Hofmann and Thomas 2000) and  
 18 LDA (Blei et al. 2003).

### Word Sense Embeddings in Sentence Embeddings

While Chap. 4 was all about sense embeddings, they are unmentioned here. One might think that they would be very useful for sentence embeddings. However, they are not as needful as one might expect. The sense of a word being used is determined by the context. Ideally, it is determined by what the context means. As a sentence embedding is a direct attempt to represent the meaning of such a context, determining the sense of each word within it is not required. Using sense embeddings instead of word embeddings is a valid extension to many of these methods. However it requires performing word sense disambiguation, which as discussed is very difficult.

---

## 5.1 Unordered and Weakly Ordered Representations

20 A model that does not take into account word order cannot perfectly capture the  
 21 meaning of a sentence. Mitchell and Lapata (2008) give the poignant examples of:

- 22 • It was not the sales manager who hit the bottle that  
 23 day, but the office worker with the serious drinking  
 24 problem.
- 25 • That day the office manager, who was drinking, hit  
 26 the problem sales worker with a bottle, but it was  
 27 not serious.

28 These two sentences have the same words, but in a different structure, resulting in  
 29 their very different meanings. In practice, however, representations which discard  
 30 word order can be quite effective.

### 5.1.1 Sums of Word Embeddings

#### SOWE is the product of the BOW with an embedding matrix

The reader may recall from Chap. 4, that a word-embedding lookup is the

LW: move this box so it is not splot across 2 pages

same as a one-hot vector product:  $C_{:,w^i} = C \hat{e}_{w^i}$ . Similar can be said for sum of word embeddings (SOWE) and bag of words (BOW). For some set of words  $\mathcal{W} = \{w_1, \dots, w_n\}$ : the BOW representation is  $B_{\mathcal{W}} = \sum_{w^i \in \mathcal{W}} \hat{e}_{w^i}$ ; the SOWE representation is  $\sum_{w^i \in \mathcal{W}} C_{w^i} = CB_{\mathcal{W}}$ . As with word-embeddings, it is immensely cheaper to calculate this via lookup and sum, rather than via matrix product; except on systems with suitable sparse matrix product tricks.

Classically, in information retrieval, documents have been represented as bags of words (BOW). That is to say a vector with length equal to the size of the vocabulary, with each position representing the count of the number of occurrences of a single word. This is much the same as a *one-hot vector* representing a word, but with every word in the sentence/document counted. The word embedding equivalent is sums of word embeddings (SOWE), and mean of word embeddings (MOWE). These methods, like BOW, lose all order information in the representation. In many cases it is possible to recover a BOW from a much lower dimensional SOWE (White et al. 2016a).

Surprisingly, these unordered methods have been found on many tasks to be extremely well performing, better than several of the more advanced techniques discussed later in this chapter. This has been noted in several works including: White et al. (2015), Ritter et al. (2015) and Wang et al. (2017). It has been suggested that this is because in English there are only a few likely ways to order any given bag of words. It has been noted that given a simple n-gram language model the original sentences can often be recovered from BOW (Horvat and Byrne 2014) and thus also from a SOWE (White et al. 2016b). Thus word-order may not in-fact be as important as one would expect in many natural language tasks, as it is in practice more proscribed than one would expect. That is to say very few sentences with the same word content, will in-practice be able to have it rearranged for a very different meaning. However, this is unsatisfying, and certainly cannot capture fine grained meaning.

The step beyond this is to encode the n-grams into a bag of words like structure. This is a bag of n-grams (BON), e.g. bag of trigrams. Each index in the vector thus represents the occurrence of an n-gram in the text. So *It is a good day today*, has the trigrams: (*It is a*),(*is a good*),(*a good day*),(*good day today*). As is obvious for all but the most pathological sentences, recovering the full sentence order from a bag of n-grams is possible even without a language model.

The natural analogy to this with word embeddings might seem to be to find n-gram embeddings by the concatenation of  $n$  word embeddings; and then to sum these. However, such a sum is less informative than it might seem. As the sum in each concatenated section is equal to the others, minus the edge words.

Instead one should train an n-gram embedding model directly. The method discussed in Chap. 4, can be adapted to use n-grams rather than words as the basic token. This was explored in detail by Li et al. (2017). Their model is based on the skip-gram word embedding method. They take as input an n-gram embedding, and attempt to predict the surrounding n-grams. This reduces to the original skip-gram method for the case of unigrams. Note that the surrounding n-grams will overlap in words (for

<sup>70</sup>  $n > 1$ ) with the input n-gram. As the overlap is not complete, this task remains  
<sup>71</sup> difficult enough to encourage useful information to be represented in the embeddings.  
<sup>72</sup> Li et al. (2017) also consider training n-gram embeddings as a bi-product of  
<sup>73</sup> text classification tasks.

### <sup>74</sup> 5.1.2 Paragraph Vector Models (Defunct)

#### Window versus Context

It is important to be clear in this section on the difference between the window and the context. The window is the words near the target word. The context (in this context) refers to the larger structure (sentence, paragraph, document) that a representation is attempting to be found for. The window is always a subset of the context. In modelling the context many windows within it will be considered (one per target word). Some works say sentence vector, document vector or paragraph vector. We say context vector as it could be any of the above. In theory it could even be a whole collection of documents.

<sup>75</sup> Le and Mikolov (2014) introduced two models for representing documents of  
<sup>76</sup> any length by using augmented word-embedding models. The models are called  
<sup>77</sup> Paragraph Vector Distributed Memory (PV-DM) model, and the Paragraph Vector  
<sup>78</sup> Distributed Bag of Words model (PV-DBOW). The name Paragraph Vector is a mis-  
<sup>79</sup> nomer, it function on texts of any length and has most often (to our knowledge)  
<sup>80</sup> been applied to documents and sentences rather than any in-between structures. The  
<sup>81</sup> CBOW and skip-gram models are extended with an additional context vector that  
<sup>82</sup> represents the current document (or other super-word structure, such as sentence or  
<sup>83</sup> paragraph). This, like the word embeddings, is initialised randomly, then trained  
<sup>84</sup> during the task. Le and Mikolov (2014) considered that the context vector itself must  
<sup>85</sup> contain useful information about the context. The effect in both cases of adding a  
<sup>86</sup> context vector is to allow the network to learn a mildly different accusal language  
<sup>87</sup> model depending on the context. To do this, the context vector would have to learn  
<sup>88</sup> a representation for the context.

#### PV Model Implementations

There is a popular third-party implementation of both the paragraph vector models, under the name doc2vec in the python gensim library (Rehůrek et al. 2010), along with many information retrieval vector models such as LDA.

<sup>89</sup> PV-DBOW is an extension of CBOW. The inputs to the model are not only the word-  
<sup>90</sup> embedding  $C_{:, w_j}$  for the words  $w^j$  from the window, but also a context-embedding  
<sup>91</sup>  $D_{:, q^k}$  for its current context (sentence, paragraph or document)  $d^k$ . The task remains

92 to predict which word was the missing word from the center of the context  $w^i$ .

$$93 \quad P(w^i | d^k, w^{i-\frac{n}{2}}, \dots, w^{i-1}, w^{i+1}, \dots, w^{i+\frac{n}{2}}) \\ 94 \quad = \text{smax} \left( WD_{:,d^k} + U \sum_{j=i+1}^{j=\frac{n}{2}} (C_{:,w^{i-j}} + C_{:,w^{i+j}}) \right) \quad (5.1)$$

96 PV-DM is the equivalent extension for skip-grams. Here the input to the model is not  
 97 only the central word, but also the context vector. Again, the task remains to predict  
 98 the other words from the window.

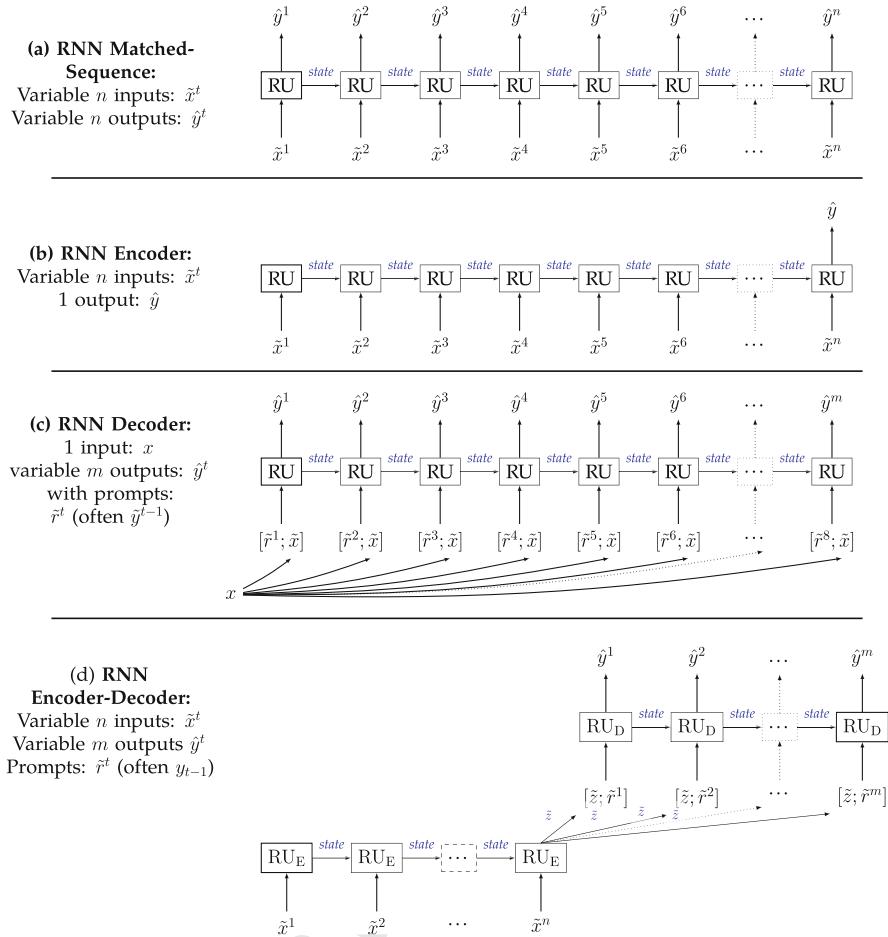
$$99 \quad P(w^j | d^k, w^i) = [\text{smax}(WD_{:,d^k} + V C_{:,w^i})]_{w_j} \quad (5.2)$$

100 The results of this work are now considered of limited validity. There were failures  
 101 to reproduce the reported results in the original evaluations which were on sentiment  
 102 analysis tasks. These were documented online by several users, including by the  
 103 second author.<sup>1</sup> A follow up paper, Mesnil et al. (2014) found that reweighed bags of  
 104 n-grams (Wang et al. 2012) out performed the paragraph vector models. Conversely,  
 105 Lau et al. (2016) found that on short text-similarity problems, with the right tuning,  
 106 the paragraph vector models could perform well; however they did not consider  
 107 the reweighed n-grams of Wang et al. (2012). On a different short text task, White  
 108 et al. (2015) found the paragraph vector models to significantly be out-performed  
 109 by SOWE, MOWE, BOW, and BOW with dimensionality reduction. This highlights  
 110 the importance of rigorous testing against a suitable baseline, on the task in question.

## 111 5.2 Sequential Models

112 The majority of this section draws on the recurrent neural networks (RNN) as dis-  
 113 cussed in Chap. 3. Every RNN learns a representation of all its input and output in  
 114 its state. We can use RNN encoders and decoders (as shown in Fig. 5.1) to generate  
 115 representations of sequences by extracting a coding layer. One can take any RNN  
 116 encoder, and select one of the hidden state layers after the final recurrent unit (RU)  
 117 that has processed the last word in the sentence. Similarly for any RNN decoder,  
 118 one can select any hidden state layer before the first recurrent unit that begins to  
 119 produce words. For an RNN encoder-decoder, this means selecting the hidden layer  
 120 from between. This was originally considered in Cho et al. (2014), when using a  
 121 machine translation RNN, to create embeddings for the translated phrases. Several  
 122 other RNNs have been used in this way since.

<sup>1</sup><https://groups.google.com/forum/#!msg/word2vec-toolkit/Q49F1rNOQRo/DoRuBoVNFb0J>.



**Fig. 5.1** The unrolled structure of an RNN for use in **a** Matched-sequence **b** Encoding, **c** Decoding and **d** Encoding-Decoding (sequence-to-sequence) problems. RU is the recurrent unit – the neural network which reoccurs at each time step. (Repeated from Fig. 3.1)

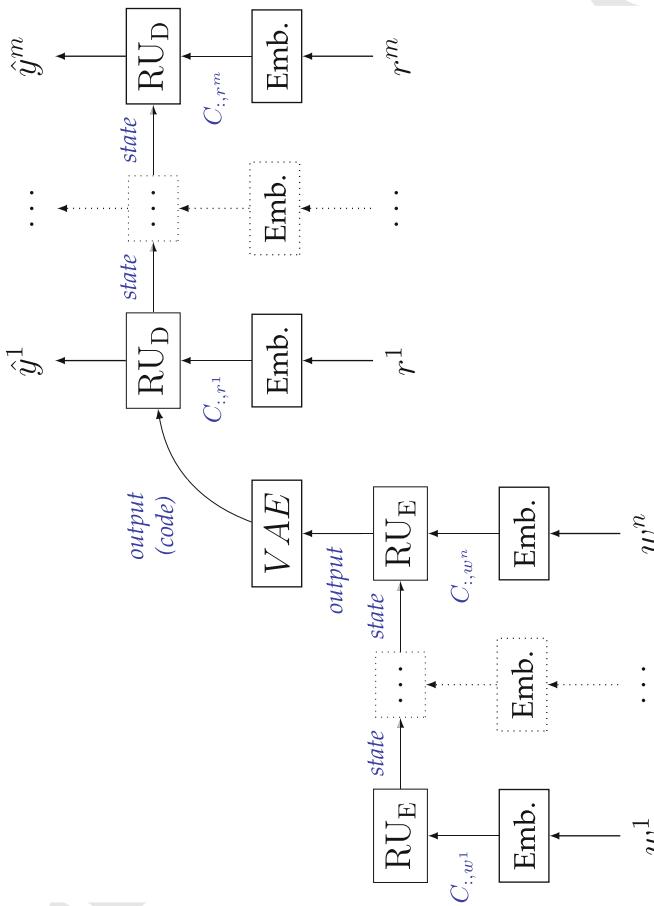
123

## 5.2.1 VAE and Encoder-Decoder

124

Bowman et al. (2016b) presents an extension on this notion, where in-between the encode and the decode stages there is a variational autoencoder (VAE). This is shown in Fig. 5.2. The variational autoencoder (Kingma and Welling 2014) has been demonstrated to have very good properties in a number of machine learning applications: they are able to work to find continuous latent variable distributions over arbitrary likelihood functions (such as in the neural network); and are very fast to train. Using the VAE, it is hoped that a better representation can be found for the sequence of words in the input and output.

LW: Shink image so it fits in normal upright orientation.



**Fig. 5.2** The VAE plus encoder-decoder of Bowman et al. (2016b). Note that during training,  $\hat{y}^i = w^i$ , as it is an autoencoder model. As is normal for encoder-decoders the prompts are the previous output (target during training, predicted during testing):  $r^i = y^{i-1}$ , with  $r^1 = y_0 = <\text{EOS}>$  being a pseudo-token marker for the end of the string. The Emb. step represents the embedding table lookup. In the diagrams for Chap. 4 we showed this as a table but just as a block here for conciseness



132 Bowman et al. (2016b) trained the network as encoder-decoder reproducing its exact  
 133 input. They found that short syntactically similar sentences were located near to each  
 134 other according to this space, further to that, because it has a decoder, it can generate  
 135 these nearby sentences, which is not possible for most sentence embedding methods.

136 Interestingly, they use the VAE output, i.e. the *code*, only as the state input to  
 137 the decoder. This is in-contrast to the encoder-decoders of Cho et al. (2014), where  
 138 the *code* was concatenated to the input at every timestep of the decoder. Bowman  
 139 et al. (2016b) investigated such a configuration, and found that it did not yield an  
 140 improvement in performance.

### 141 5.2.2 Skip-Thought

142 Kiros et al. (2015) draws inspiration from the works on acausal language modelling,  
 143 to attempt to predict the previous and next sentence. Like in the acausal language  
 144 modelling methods, this task is not the true goal. Their true goal is to capture a good  
 145 representation of the current sentence. As shown in Fig. 5.3 they use an encoder-  
 146 decoder RNN, with two decoder parts. One decoder is to produce the previous sen-  
 147 tence. The encoder part takes as its input the current sentence, and produces as its  
 148 output the code, which is input to the decoders. The other decoder is to produce the  
 149 next sentence. As described in Sect. 3.2.3, the prompt used for the decoders includes  
 150 the previous word, concatenated to the code (from the encoder output).

151 That output code is the representation of the sentence.

## 152 5.3 Structured Models

### Parsers

There are many high-quality parsing libraries available. The most well known is the Stanford CoreNLP library (Manning et al. 2014) for java. It has an interactive web-demo at <http://corenlp.run/>, which was used to produce Figs. 5.4 and 5.5. NLTK (Bird et al. 2009) contains several different parsers, including a binding to CoreNLP parsers. The newer spaCy library (Honnibal et al. 2015) for python, presently only features a dependency parser.

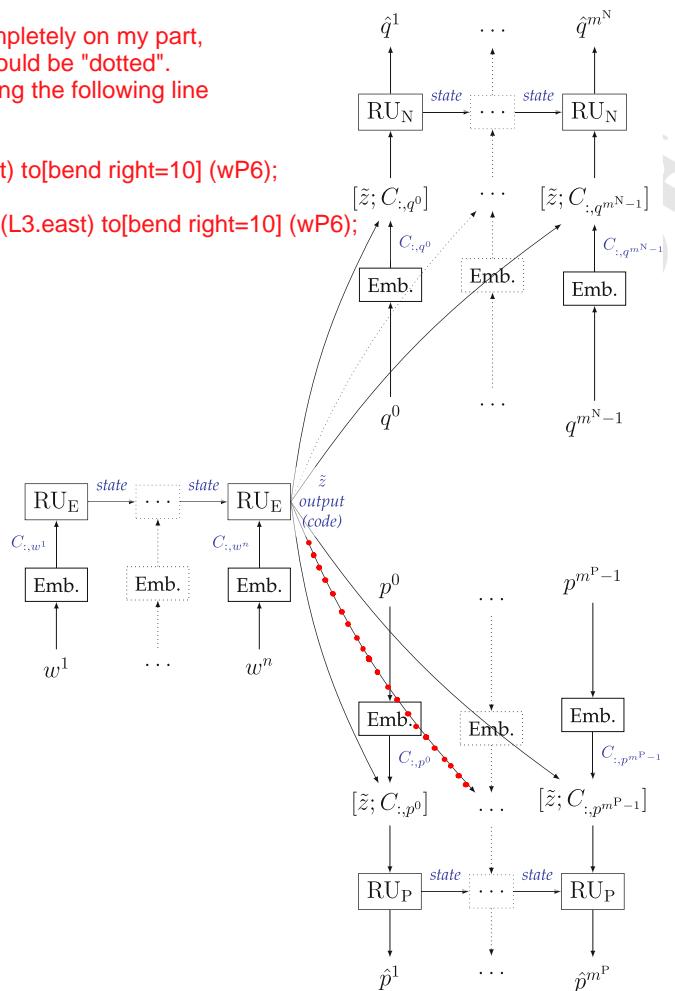
153 The sequential models are limited to process the information as a series of time-  
 154 steps one after the other. They processes sentences as ordered lists of words. However,  
 155 the actual structure of a natural language is not so simple. Linguists tend to break  
 156 sentences down into a tree structure. This is referred to as parsing. The two most  
 157 common forms are constituency parse trees, and dependency parse trees. Examples  
 158 of each are shown in Figs. 5.4 and 5.5. It is beyond the scope of this book to explain the  
 159 precise meaning of these trees, and how to find them. The essence is that these trees

LW: It is a mistake completely on my part,  
but the marked line should be "dotted".  
This is done by changing the following line  
in the TeX:

```
\path[draw, ->] (L3.east) to[bend right=10] (wP6);
```

To

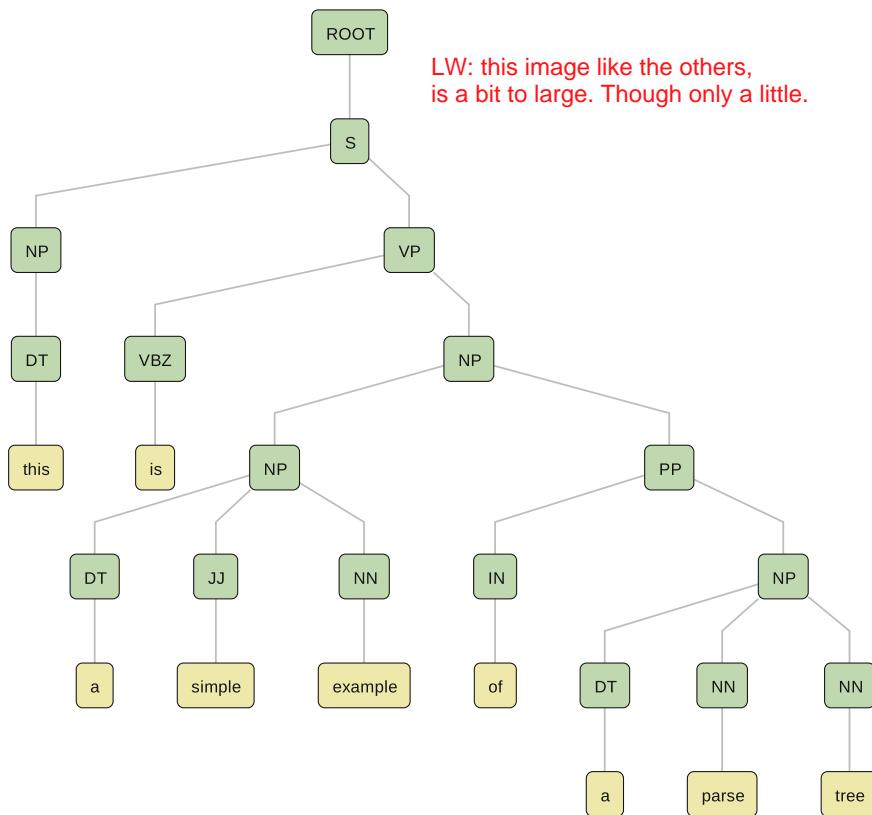
```
\path[draw, ->,dotted] (L3.east) to[bend right=10] (wP6);
```



**Fig. 5.3** The skip-thought model (Kiros et al. 2015). Note that for the next and previous sentences respectively the outputs are  $\hat{q}^i$  and  $\hat{p}^i$ , and the prompts are  $q^{i-1}$  and  $p^{i-1}$ . As there is no intent to use the decoders after training, there is no need to worry about providing an evaluation-time prompt, so the prompt is always the previous word.  $p^0 = p^{m^P} = q^0 = q^{m^Q} = <\text{EOS}>$  being a pseudo-token marker for the end of the string. The input words are  $w^i$ , which come from the current sentence. the Emb. steps represents the look-up of the embedding for the word.

160 represent the structure of the sentence, according to how linguists believe sentences  
161 are processed by humans.

162 The constituency parse breaks the sentence down into parts such as noun phrase  
163 (NP) and verb phrase (VP), which are in turn broken down into phrases, or (POS  
164 tagged) words. The constituency parse is well thought-of as a hierarchical breakdown  
165 of a sentence into its parts. Conversely, a dependency parse is better thought of as a set



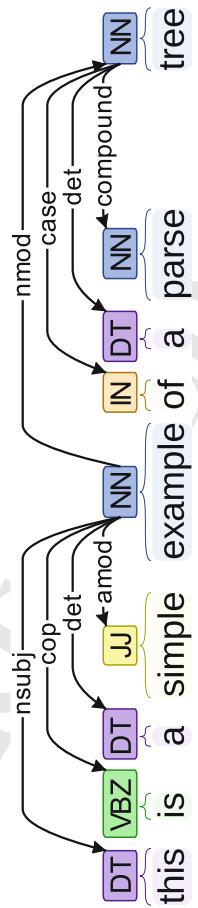
**Fig. 5.4** A constituency parse tree for the sentence: This is a simple example of a parse tree. In this diagram the leaf nodes are the input words, their immediate parents are their POS tags, and the other nodes with multiple children represent sub-phrases of the sentence, for example NP is a Noun Phrase

166 of binary relations between head-terms and their dependent terms. These structures  
 167 are well linguistically motivated, so it makes sense to use them in the processing of  
 168 natural language.

169 We refer here to models incorporating tree (or graph) structures as structural  
 170 models. Particular variations have their own names, such as recursive neural networks  
 171 (RvNN), and recursive autoencoders (RAE). We use the term structural model as an  
 172 all encompassing term, and minimise the use of the easily misread terms: recursive  
 173 versus recurrent neural networks. A sequential model (an RNN) is a particular case  
 174 of a structural model, just as a linked list is a particular type of tree. However, we  
 175 will exclude sequential models from this discussion except where marked.

176 The initial work on structural models was done in the thesis of Socher (2014). It  
 177 builds on the work of Goller and Kuchler (1996) and Pollack (1990), which present  
 178 back-propagation through structure. Back-propagation can be applied to networks

LW: Shrink this image such that it fits upright



**Fig. 5.5** A dependency parse tree for the sentence This is a simple example of a parse tree. This flattened view may be misleading, example is at the peak of the tree, with direct children being: this, is, a, simple, and tree. tree has direct children being: of, a, and parse

179 of any structure, as the chain-rule can be applied to any differentiable equation to  
180 find its derivative. Structured networks, like all other networks, are formed by the  
181 composition of differentiable functions, so are differentiable. In a normal network  
182 the same composition of functions is used for all input cases, whereas in a structured  
183 network it is allowed to vary based on the inputs. This means that structuring a  
184 network according to its parse tree is possible.

### Machine learning frameworks for structural models

Structural networks cannot be easily defined in most static neural network libraries, such as TensorFlow. These implementations function by defining a single computational graph that is used to process each training/test case. The same graph is used for each input. By definition, the structure of the network differs from training/test case to training/test case. Technically the same problems apply to RNNs, as each case can have a different number of inputs. This is normally worked around by defining the network graph for the longest input to be considered, then padding all the inputs to this length, and ensuring that the padding does not interfere with the gradient updates. The equivalent tricks for structured networks are significantly more complex. The exception to this is of-course via dynamic components to the static frameworks (which TensorFlow and other such frameworks certainly do have). Even in a dynamic framework it remains a non-trivial task to implement these networks.

### Implementing Back-propagation through structure

Conceptually, back-propagation through structure is not significantly more complex than back-propagation through time. However, in practice it is a very difficult algorithm to get right. It is very important to test for correctness using gradient checks, as it is easy to make a mistake and end-up with models that seem to work ok, but are actually crippled due to some mistake in the coding. Unfolding recursive autoencoders are particularly difficult, as the gradient must be propagated from all leaves. And output interior nodes cannot have their gradients calculated until the gradients of their children are calculated. The solution to this is to process the node gradient calculation using a priority queue, where the priority is set by the depth of the node. Thus ensuring that all children are processed before their parents.

#### 5.3.1 Constituency Parse Tree (Binary)

185 Tree structured networks work by applying a recursive unit (which we will call RV)  
186 function across pairs (or other groups) of the representations of the lower levels, to  
187 produce a combined representation. The network structure for an input of binary tree  
188

structured text is itself a binary tree of RVs. Each RV (i.e. node in the graph) can be defined by the composition function:

$$f^{\text{RV}}(\tilde{u}, \tilde{v}) = \varphi \left( [S \ R] \begin{bmatrix} \tilde{u} \\ \tilde{v} \end{bmatrix} + \tilde{b} \right) \quad (5.3)$$

$$= \varphi \left( S\tilde{u} + R\tilde{v} + \tilde{b} \right) \quad (5.4)$$

where  $\tilde{u}$  and  $\tilde{v}$  are the left and right substructures embeddings (word embeddings at the leaf node level), and  $S$  and  $R$  are the matrices defining how the left and right children's representations are to be combined.

This is a useful form as all constituency parse trees can be converted into binary parse trees, via left-factoring or right factoring (adding new nodes to the left or right to take some of the children). This is sometimes called binarization, or putting them into Chomsky normal form. This form of structured network has been used in many words, including (Socher et al. 2010, 2011a,b,c; Zhang et al. 2014). Notice that  $S$  and  $R$  matrices are shared for all RVs, so all substructures are composed in the same way, based only on whether they are on the left, or the right.

### 5.3.2 Dependency Tree

The dependency tree is the other commonly considered parse-tree. Structured networks based upon the dependency tree have been used by Socher et al. (2014), Iyyer et al. (2014a,b). In these works rather than using composition matrix for left-child and right-child, the composition matrix varies depending on the type of relationship of between the head word and its child. Each dependency relationship type has its own composition matrix. That is to say there are distinct composition matrices for each of nsubj, det, nmod, case etc. This allows for multiple inputs to a single head node to be distinguished by their relationship, rather than their order. This is important for networks using a dependency parse tree structure as the relationship is significant, and the structure allows a node to have any number of inputs.

#### Extended example

The full example for the  $f^{\text{RV}}(9)$  from Eq. 5.5 is:

$$\begin{aligned} f^{\text{RV}}(9) &= \varphi(W^{\text{head}} C_{:, \text{tree}} \\ &+ W^{\text{compound}} (W^{\text{head}} C_{:, \text{parse}} + \tilde{b}) \\ &+ W^{\text{det}} (W^{\text{head}} C_{:, \text{a}} + \tilde{b}) \\ &+ W^{\text{case}} (W^{\text{head}} C_{:, \text{of}} + \tilde{b}) \\ &+ \tilde{b}) \end{aligned}$$

LW: Realign this, to align with the mark

This in turn would be composed as part of  $f^{\text{RV}}(5)$  for the whole tree headed by  $w^5 = \text{example}$ . The output of each RV is a representation of that substructure.

### No gates LW: Remove line break

#### No long-term memory

We note that a limitation of most structural models, compared to the sequential RNNs, is their lack of explicit gating on memory (e.g. as in GRU and LSTM). Any given path down a tree can be looked at as a simple RNN comprised only of basic recurrent units. However, these paths are much shorter (being the logarithm of) than the full sequential length of the sentence, which offsets the need for such gating. Recall that the gating is to provide the longer short term memory.

215 Consider a function  $\pi(i, j)$  which returns the relationship between the head word  
 216 at position  $i$  and the child word at position  $j$ . For example, using the tree shown in  
 217 Fig. 5.5, which has  $w^8 = \text{parse}$  and  $w^9 = \text{tree}$  then  $\pi(8, 9) = \text{compound}$ . This  
 218 is used to define the composed representation for each RV:

$$219 \quad f^{\text{RV}}(i) = \varphi \left( W^{\text{head}} C_{:, w^i} + \sum_{j \in \text{children}(i)} W_{\pi(i, j)} f_{\text{RV}}(j) + \tilde{b} \right) \quad (5.5)$$

220 Here  $C_{:, w^i}$  is the word embedding for  $w^i$ , and  $W_{\text{head}}$  encodes the contribution  
 221 of the headword to the composed representation. Similarly,  $W_{\pi(i, j)}$  encodes the  
 222 contribution of the child words. Note that the terminal case is just  $f_{\text{RV}}(i) =$   
 223  $\varphi(W^{\text{head}} C_{:, w^i} + \tilde{b})$  when a node  $i$  has no children. This use of the relationship  
 224 to determine the composition matrix, increases both the networks expressiveness,  
 225 and also handles the non-binary nature of dependency trees.

226 A similar technique could be applied to constituency parse trees. This would be  
 227 using the part of speech (e.g. VBZ, NN) and phrase tags (e.g. NP, VP) for the sub-  
 228 structures to choose the weight matrix. This would, however, lose the word-order  
 229 information when multiple inputs have the same tag. This would be the case, for  
 230 example, in the right-most branch shown in Fig. 5.4, where both `parse` and `tree`  
 231 have the NN POS tag, and thus using only the tags, rather than the order would  
 232 leave `parse tree` indistinguishable from `tree parse`. This is not a problem  
 233 for the dependency parse, as word relationships unambiguously correspond to the  
 234 role in the phrase's meaning. As such, allowing the dependency relationship to define  
 235 the mathematical relationship, as encoded in the composition matrix, only enhances  
 236 expressibility.

237 For even greater capacity for the inputs to control the composition, would be to  
 238 allow every word be composed in a different way. This can be done by giving the  
 239 child nodes there own composition matrices, to go with there embedding vectors.  
 240 The composition matrices encode the relationship, and the operation done in the

composition. So not only is the representation of the (sub)phrase determined by a relationship between its constituents (as represented by their embeddings), but the nature of that relationship (as represented by the matrix) is also determined by those same constituents. In this approach at the leaf-nodes, every word not only has a word vector, but also a word matrix. This is discussed in Sect. 5.4.

### 5.3.3 Parsing

The initial work for both contingency tree structured networks (Socher et al. 2010) and for dependency tree structured networks (Stenetorp 2013) was on the creation of parsers. This is actually rather different to the works that followed. In other works the structure is provided as part of the input (and is found during preprocessing). Whereas a parser must induce the structure of the network, from the unstructured input text. This is simpler for contingency parsing, than for dependency parsing.

#### The finer detail of parsing

Parsing is one of the most well studied problems in computational linguistics. Presented here is only the highest level overview. For more details on this, we recommend consulting the source materials. Ideally, with reference to a good traditional (that is to say non-neural network based) NLP textbook, such as: Manning et al. (1999).

#### Getting the Embeddings out of the Parser

The implementation of Socher et al. (2010), is publicly available. However, it does not export embeddings. It is nested deep inside the Stanford Parser, and thus accessing the embeddings is not at all trivial.

When creating a binary contingency parse tree, any pair of nodes can only be merged if they are adjacent. The process described by Socher et al. (2010), is to consider which nodes are to be composed into a higher level structure each in turn. For each pair of adjacent nodes, an RV can be applied to get a merged representation. A linear scoring function is also learned, that takes a merged representation and determines how good it was. This is trained such that correct merges score highly. Hinge loss is employed for this purpose. The Hinge loss function works on similar principles to negative sampling (see the motivation given in Sect. 4.4.2). Hinge loss is used to cause the merges that occur in the training set to score higher than those that do not. To perform the parse, nodes are merged; replacing them with their composed representation; and the new adjacent pairing score is then recomputed. Socher et al. (2010) considered both greedy, and dynamic programming search to determine the order of composition, as well as a number of variants to add additional information to the process. The dependency tree parser extends beyond this method.

Dependency trees can have child-nodes that do not correspond to adjacent words (non-projective language). This means that the parser must consider that any (unlinked) node be linked to any other node. Traditional transition-based dependency parsers function by iteratively predicting links (transitions) to add to the structure based on its current state. Stenetorp (2013) observed that a composed representation similar to Eq. 5.3, was an ideal input to a softmax classifier that would predict the next link to make. Conversely, the representation that is suitable for predicting the next link to make, is itself a composed representation. Note, that Stenetorp (2013) uses the same matrices for all relationships (unlike the works discussed in Sect. 5.3.2). This is required, as the relationships must be determined from the links made, and thus are not available before the parse. Bowman et al. (2016a), presents a work an an extension of the same principles, which combines the parsing step with the processing of the data to accomplish some task, in their case detecting entailment.

#### 5.3.4 Recursive Autoencoders

##### Application to image retrieval

An interesting application of structured networks was shown in Socher et al. (2014). A dependency tree structured network was trained on a language modelling task (not as a recursive autoencoder, although that would also have been a valid option). Then, separately a convolutional neural network was trained to produce a vector output of the same dimensionality – an image embedding – such that its distance to its caption’s composed vector representation was minimised. The effect of this was that images and their captions are projected to a common vector space. This allowed for smart image retrieval, from descriptions, by having a set of all images, and storing their embedding representations. Then for any query, the sentence embedding can be found and the vector space of images can be searched for the nearest. The reverse is not generally as useful, as one can’t reasonably store all possible captions describing an image, so as to be able to search for the best one for a user provided image. This process of linking a sequence representation to an image embedding is not restricted to structured networks, and can be done with any of the representation methods discussed in this chapter. Further, as discussed in Sect. 4.6 it can also be done using pretrained embedding on both sides through (kernel) CCA.

##### Unfolding RAE implementation

The implementation, and a pretrained model, of the unfolding recursive autoencoder of Socher et al. (2011a) is available online at <https://tinyurl.com/URAE2011>. It is easy to use as a command-line Matlab script to generate embeddings.

281 Recursive autoencoders are autoencoders, just as the autoencoder discussed in  
282 Sect. 2.5.2, they reproduce their input. It should be noted that unlike the encoder-  
283 decoder RNN discussed in Sect. 5.2.1, they cannot be trivially used to generate nat-  
284 ural language from an arbitrary embeddings, as they require the knowledge of the  
285 tree structure to unfold into. Solving this would be the inverse problem of parsing  
286 (discussed in Sect. 5.3.3).

287 The models presented in Socher et al. (2011a) and Iyyer et al. (2014a) are unfolding  
288 recursive autoencoders. In these models an identical inverse tree is defined above the  
289 highest node. The loss function is the sum of the errors at the leaves, i.e. the distance  
290 in vector space between the reconstructed words embeddings and the input word-  
291 embeddings. This was based on a simpler earlier model: the normal (that is to say,  
292 not unfolding) recursive autoencoder.

293 The normal recursive autoencoder, as used in Socher et al. (2011c) and Zhang  
294 et al. (2014) only performs the unfolding for a single node at a time during training.  
295 That means that it assesses how well each merge can individually be reconstructed,  
296 not the success of the overall reconstruction. This per merge reconstruction has a  
297 loss function based on the difference between the reconstructed embeddings and the  
298 inputs embeddings. Note that those inputs/reconstructions are not word embeddings:  
299 they are the learned merged representations, except when the inputs happen to be  
300 leaf node. This single unfold loss covers the auto-encoding nature of each merge;  
301 but does not give any direct assurances of the auto-encoding nature of the whole  
302 structure. However, it should be noted that while it is not trained for, the reconstruction  
303 components (that during training are applied only at nodes) can nevertheless be  
304 applied recursively from the top layer, to allow for full reconstruction.

### 305 5.3.4.1 Semi-supervised

306 In the case of all these autoencoders, except (Iyyer et al. 2014a), a second source of  
307 information is also used to calculate the loss during training. The networks are being  
308 simultaneously trained to perform a task, and to regenerate their input. This is often  
309 considered as semi-supervised learning, as unlabelled data can be used to train the  
310 auto-encoding part (unsupervised) gaining a good representation, and the labelled  
311 data can be used to train the task output part (supervised) making that representation  
312 useful for the task. This is done by imposing an additional loss function onto the  
313 output of the central/top node.

- 314 • In Socher et al. (2011c) this was for sentiment analysis.  
315 • In Socher et al. (2011a) this was for paraphrase detection.  
316 • In Zhang et al. (2014) this was the distance between embeddings of equivalent  
317 translated phrases of two RAEs for different languages.

318 The reconstruction loss and the supervised loss can be summed, optimised in alter-  
319 nating sequences, or the reconstructed loss can be optimised first, then the labelled  
320 data used for fine-tuning.

### Sequential models are often preferred to structural models

Sequential (RNN) models are much more heavily researched than structural models. They have better software libraries, are easier to implement, and have more known “tricks” (like gates and attention). In theory it is possible for a sequential model (with sufficiently deep and wide RNNs) to internalise the connections that a structural model would possess. While structural models are theoretically nicer from a linguistics standpoint, pragmatically they are the last resort in modelling. When attempting to find a useful representation of a sentence for a task, one should first try a sum of word embeddings with a simple network on-top, then a sequential model (based on LSTM or GRU), and only if these fail then try a structured model. Arguably before using any neural network approach at all, one should eliminate bag of words, bag of n-grams, the dimensionality reduced version of those bags, and also eliminate LSI and LDA as inputs for the task.

321

## 5.4 Matrix Vector Models

322

### 5.4.1 Structured Matrix Vector Model

323

Socher et al. (2012) proposed that each node in the graph should define not only a vector embedding, but a matrix defining how it was to be combined with other nodes. That is to say, each word and each phrase has both an embedding, and a composition matrix.

327

Consider this for binary constituency parse trees. The composition function is as follows:

329

$$f^{RV}(\tilde{u}, \tilde{v}, U, V) = \varphi([S \ R][U\tilde{v}; V\tilde{u}] + \tilde{b}) \quad (5.6)$$

330

$$= \varphi(SU\tilde{v} + RV\tilde{u} + \tilde{b}) \quad (5.7)$$

331

$$F^{RV}(U, V) = W[U; V] = W_l U + W_r V \quad (5.8)$$

333

$f^{RV}$  gives the composed embedding, and  $F^{RV}$  gives the composing matrix. The  $S$  and  $R$  represent the left and right composition matrix components that are the same for all nodes (regardless of content). The  $U$  and  $V$  represent the per word/node child composition matrix components. We note that  $S$  and  $R$  could, in theory, be rolled in to  $U$  and  $R$  as part of the learning. The  $\tilde{u}$  and  $\tilde{v}$  represent the per word/node children embeddings, and  $W$  represents the method for merging two composition matrices.

339

We note that one can define increasingly complex and powerful structured networks along these lines; though one does run the risk of very long training times and of over-fitting.

### 342 5.4.2 Sequential Matrix Vector Model

343 A similar approach, of capturing a per word matrix, was used on a sequential model by  
 344 Wang et al. (2017). While sequential models are a special case of structured models,  
 345 it should be noted that unlike the structured models discussed prior, this matrix vector  
 346 RNN features a gated memory. This matrix-vector RNN is an extension of the GRU  
 347 discussed in Chap. 3, but without a reset gate.

348 In this sequential model, advancing a time step, is to perform a composition.  
 349 This composition is for between the input word and the (previous) state. Rather than  
 350 directly between two nodes in the network as in the structural case. It should be  
 351 understood that composing with the state is not the same as composing the current  
 352 input with the previous input. But rather as composing the current input with all  
 353 previous inputs (though not equally).

354 As depicted in Fig. 5.6 each word,  $w^t$  is represented by a word embedding  $\tilde{x}^t$  and  
 355 matrix:  $\tilde{X}^{w^t}$ , these are the inputs at each time step. The network outputs and states  
 356 are the composed embedding  $\hat{y}^t$  and matrix  $Y^t$ .

**Remember:**

The product of a matrix with a concatenated vector, is the same as the sum of  
 the two blocks of that matrix each multiplied by the blocks of that vector.

$$357 \quad h^t = \tanh \left( W^h [x^t; \hat{y}^{t-1}] + \tilde{b}^h \right) \quad (5.9)$$

$$358 \quad z^t = \sigma \left( Y^{t-1} x^t + X^t \hat{y}^{t-1} + \tilde{b}^z \right) \quad (5.10)$$

$$359 \quad \hat{y}^t = z^t \odot h^t + (1 - z^t) \odot \hat{y}^{t-1} \quad (5.11)$$

$$360 \quad Y^t = \tanh \left( W^Y [Y^{t-1}; X^t] + \tilde{b}^Y \right) \quad (5.12)$$

362 The matrices  $W^h$ ,  $W^Y$  and the biases  $\tilde{b}^h$ ,  $\tilde{b}^z$ ,  $\tilde{b}^Y$  are shared across all time  
 363 steps/compositions.  $W^Y$  controls how the next state-composition  $Y^t$  matrix is gen-  
 364 erated from its previous value and the input composition matrix,  $X^t$ ;  $W^h$  similarly  
 365 controls the value of the candidate state-embedding  $h^t$ .

366  $h^t$  is the candidate composed embedding (to be output/used as state).  $z^t$  is the  
 367 update gate, it controls how much of the actual composed embedding ( $\hat{y}^t$ ) comes  
 368 from the candidate  $h^t$  and how much comes from the previous value ( $\hat{y}^{t-1}$ ). The  
 369 composition matrix  $Y^t$  (which is also part of the state/output) is not gated.

370 Notice, that the state composition matrix  $Y^{t-1}$  is only used to control the gate  $z^t$ ,  
 371 not to directly affect the candidate composed embedding  $h^t$ . Indeed, in fact one can  
 372 note that all similarity to the structural method of Socher et al. (2012) is applied in  
 373 the gate  $z^t$ . The method for calculating  $h^t$  is similar to that of a normal RU.

LW: Shrink this image such that it fits upright

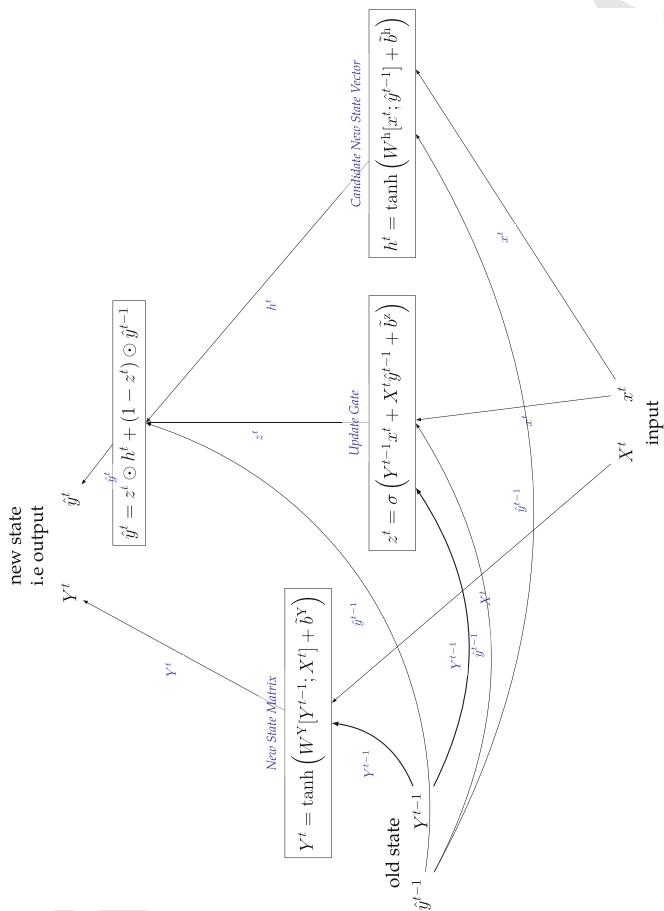


Fig. 5.6 A Matrix Vector recurrent unit

374 The work of Wang et al. (2017), was targeting short phrases. This likely explains  
375 the reason for not needing a forget gates. The extension is obvious, and may be  
376 beneficial when applying this method to sentences.

## 377 5.5 Conclusion, on Compositionality

378 It is tempting to think of the structured models as compositional, and the sequential  
379 models as non-compositional. However, this is incorrect.

380 The compositional nature of the structured models is obvious: the vector for a  
381 phrase is composed from the vectors of the words that the phrase is formed from.

382 Sequential models are able to learn the structures. For example, learning that  
383 a word from  $n$  time steps ago is to be remembered in the RNN state, to then be  
384 optimally combined with the current word, in the determination of the next state.  
385 This indirectly allows the same compositionality as the structured models. It has been  
386 shown that sequential models are indeed in-practice able to learn such relationships  
387 between words (White et al. 2017). More generally as almost all aspects of language  
388 have some degree of compositionality, and sequential models work very well on  
389 most language tasks, this implicitly shows that they have sufficient representational  
390 capacity to learn sufficient degrees of compositional processing to accomplish these  
391 tasks.

392 In fact, it has been suggested that even some unordered models such as sum of  
393 word embeddings are able to capture some of what would be thought of as compo-  
394 sitional information. Ritter et al. (2015) devised a small corpus of short sentences  
395 describing containing relationships between the locations of objects. The task and  
396 dataset was constructed such that a model must understand some compositionality,  
397 to be able to classify which relationships were described. Ritter et al. (2015)  
398 tested several sentence representations as the input to a naïve Bayes classifier being  
399 trained to predict the relationship. They found that when using sums of high-quality  
400 word embeddings as the input, the accuracy not only exceeded the baseline, but even  
401 exceeded that from using representation from a structural model. This suggests that  
402 a surprising amount of compositional information is being captured into the embed-  
403 dings; which allows simple addition to be used as a composition rule. Though it  
404 being ignorant of word order does mean it certainly couldn't be doing so perfectly,  
405 however the presence of other words my be surprisingly effective hinting at the word  
406 order (White et al. 2016b), thus allow for more apparently compositional knowledge  
407 to be encoded than is expected.

408 To conclude, the compositionality capacity of many models is not as clear cut as  
409 it may initially seem. Further to that the requirement for a particular task to actually  
410 handle compositional reasoning is also not always present, or at least not always a  
411 significant factor in practical applications. We have discussed many models in this  
section, and their complexity varies significantly. They range from the very simple

412 sum of word embeddings all the way to the the structured matrix models, which are  
 413 some of the more complicated neural networks ever proposed.

414

---

## 415 References

- 416 Bird, Steven, Ewan Klein, and Edward Loper. 2009. Natural language processing with Python.  
 417 O'Reilly Media, Inc.
- 418 Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *The Journal  
 419 of Machine Learning Research* 3: 993–1022.
- 420 Bowman, Samuel R., Jon Gauthier, Abhinav Rastogi, Raghav Gupta, Christopher D. Manning,  
 421 and Christopher Potts. 2016a. A fast unified model for parsing and sentence understanding.  
 422 [arXiv:1603.06021](https://arxiv.org/abs/1603.06021).
- 423 Bowman, Samuel R., Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Jozefowicz, and Samy  
 424 Bengio. 2016b. Generating sentences from a continuous space. In *International conference on  
 425 learning representations (ICLR) Workshop*.
- 426 Cho, Kyunghyun, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares,  
 427 Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-  
 428 decoder for statistical machine translation. In *Proceedings of the 2014 conference on empirical  
 429 methods in natural language processing (EMNLP)*, 1724–1734. Doha, Qatar: Association for  
 430 Computational Linguistics.
- 431 Dumais, Susan T., George W. Furnas, Thomas K. Landauer, Scott Deerwester, and Richard Harsh-  
 432 man. 1988. Using latent semantic analysis to improve access to textual information. In *Proceed-  
 433 ings of the SIGCHI conference on Human factors in computing systems*, 281–285. ACM.
- 434 Goller, Christoph and Andreas Kuchler. 1996. Learning task-dependent distributed representations  
 435 by back propagation through structure. In *IEEE international conference on neural networks,  
 436 1996*, vol. 1, 347–352. IEEE.
- 437 Hofmann, Thomas. 2000. Learning the similarity of documents: An information geometric approach  
 438 to document retrieval and categorization. In *Advances in neural information processing systems*,  
 439 914–920.
- 440 Honnibal, Matthew and Mark Johnson. 2015. An improved non-monotonic transition system for  
 441 dependency parsing. In *Proceedings of the 2015 conference on empirical methods in natural  
 442 language processing*, 1373–1378. Lisbon, Portugal: Association for Computational Linguistics.
- 443 Horvat, Matic and William Byrne. 2014. A graph-based approach to string regeneration. In *EACL*,  
 444 85–95.
- 445 Iyyer, Mohit, Jordan Boyd-Graber, and Hal Daumé III. 2014a. Generating sentences from semantic  
 446 vector space representations. In *NIPS workshop on learning semantics*.
- 447 Iyyer, Mohit, Jordan Boyd-Graber, Leonardo Claudino, Richard Socher, and Hal Daumé III. 2014b.  
 448 A neural network for factoid question answering over para-graphs. In *Proceedings of the 2014  
 449 conference on empirical methods in natural language processing (EMNLP)*, 633–644.
- 450 Kingma, D. P. and M. Welling. 2014. Auto-encoding variational bayes. In *The international con-  
 451 ference on learning representations (ICLR)*. [arXiv:1312.6114](https://arxiv.org/abs/1312.6114) [stat.ML].
- 452 Kiros, Ryan, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba,  
 453 Raquel Urtasun, and Sanja Fidler. 2015. Skip-thought vectors. In *CorR*. [arXiv:1506.06726](https://arxiv.org/abs/1506.06726).
- 454 Lau, Jey Han, and Timothy Baldwin. 2016. An empirical evaluation of doc2vec with practical  
 455 insights into document embedding generation. In *ACL*, 78.
- 456 Le, Quoc and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In  
 457 *Proceedings of the 31st international conference on machine learning (ICML-14)*, 1188–1196.

- 458 Li, Bofang, Tao Liu, Zhe Zhao, Puwei Wang, and Xiaoyong Du. 2017. Neural bag-of-ngrams. In  
459 *AAAI*, 3067–3074.
- 460 Manning, C.D. and H. Schütze. 1999. Foundations of statistical natural language processing. Cam-  
461 bridge: MIT Press. ISBN: 9780262133609.
- 462 Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David  
463 McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for*  
464 *computational linguistics (ACL) system demonstrations*, 55–60.
- 465 Mesnil, Grégoire, Tomas Mikolov, Marc'Aurelio Ranzato, and Yoshua Bengio. 2014. Ensem-  
466 ble of generative and discriminative techniques for sentiment analysis of movie reviews.  
467 [arXiv:1412.5335](https://arxiv.org/abs/1412.5335).
- 468 Mitchell, Jeff and Mirella Lapata. 2008. Vector-based models of semantic composition. In *ACL*,  
469 236–244.
- 470 Pollack, Jordan B. 1990. Recursive distributed representations. *Artificial Intelligence*, 46 (1): 77–  
471 105. ISSN: 0004-3702. [https://doi.org/10.1016/0004-3702\(90\)90005-K](https://doi.org/10.1016/0004-3702(90)90005-K).
- 472 Rehůrek, Radim and Petr Sojka. 2010. Software framework for topic modelling with large corpora.  
473 English. In *Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks*,  
474 45–50. Valletta, Malta: ELRA. <http://is.muni.cz/publication/884893/en>
- 475 Ritter, Samuel, Cotie Long, Denis Paperno, Marco Baroni, Matthew Botvinick, and Adele  
476 Goldberg. 2015. Leveraging preposition ambiguity to assess compositional distributional models  
477 of semantics. In *The fourth joint conference on lexical and computational semantics*.
- 478 Socher, Richard. 2014. Recursive deep learning for natural language processing and computer  
479 vision. Ph.D. thesis. Stanford University.
- 480 Socher, Richard, Christopher D. Manning, and Andrew Y. Ng. 2010. Learning continuous phrase  
481 representations and syntactic parsing with recursive neural networks. In *Proceedings of the NIPS-  
482 2010 deep learning and unsupervised feature learning workshop*, 1–9.
- 483 Socher, Richard, Eric H. Huang, Jeffrey Pennington, Andrew Y. Ng, and Christopher D. Man-  
484 ning. 2011a. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In  
485 *Advances in neural information processing systems*, 24.
- 486 Socher, Richard, Cliff C Lin, Chris Manning, and Andrew Y Ng. 2011b. Parsing natural scenes  
487 and natural language with recursive neural networks. In *Proceedings of the 28th international  
488 conference on machine learning (ICML-11)*, 129–136.
- 489 Socher, Richard, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, and Christopher D.  
490 Manning. 2011c. Semi-supervised recursive autoencoders for predicting sentiment distributions.  
491 In *Proceedings of the 2011 conference on empirical methods in natural language processing  
(EMNLP)*.
- 493 Socher, Richard, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. 2012. Semantic compo-  
494 sitionality through recursive matrix-vector spaces. In *Proceedings of the 2012 joint conference on  
495 empirical methods in natural language processing and computational natural language learning*,  
496 1201–1211. Association for Computational Linguistics.
- 497 Socher, Richard, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, and Y.Ng Andrew. 2014.  
498 Grounded compositional semantics for finding and describing images with sentences. *Transac-  
499 tions of the Association for Computational Linguistics* 2: 207–218.
- 500 Stenetorp, Pontus. 2013. Transition-based dependency parsing using recursive neural networks.  
501 In *Deep learning workshop at the 2013. Conference on neural information processing systems  
(NIPS)*. Nevada, USA: Lake Tahoe.
- 503 Wang, Sida and Christopher D. Manning. 2012. Baselines and bigrams: Simple, good sentiment  
504 and topic classification. In *Proceedings of the 50th annual meeting of the association for compu-  
505 tational linguistics: Short papers*, vol. 2, 90–94. Association for Computational Linguistics.
- 506 Wang, Rui, Wei Liu, and Chris McDonald. 2017. A matrix-vector recurrent unit model for capturing  
507 compositional semantics in phrase embeddings. In *International conference on information and  
508 knowledge management*.

- 509 White, Lyndon, Roberto Togneri, Wei Liu, and Mohammed Bennamoun. 2015. How well sentence  
510 embeddings capture meaning. In *Proceedings of the 20th Australasian document computing*  
511 *symposium. ADCS '15*, 9:1–9:8. Parramatta, NSW, Australia: ACM. ISBN: 978-1-4503-4040-3,  
512 <https://doi.org/10.1145/2838931.2838932>.
- 513 White, Lyndon, Roberto Togneri, Wei Liu, and Mohammed Bennamoun. 2016a. Generating bags  
514 of words from the sums of their word embeddings. In *17th international conference on intelligent*  
515 *text processing and computational linguistics (CICLing)*.
- 516 White, Lyndon, Roberto Togneri, Wei Liu, and Mohammed Bennamoun. 2016b. Modelling sentence  
517 generation from sum of word embedding vectors as a mixed integer programming problem. In  
518 *IEEE international conference on data mining: High dimensional data mining workshop (ICDM:*  
519 *HDM)*. <https://doi.org/10.1109/ICDMW.2016.0113>.
- 520 White, L., R. Togneri, W. Liu, and M. Bennamoun. 2017. Learning distributions of meant color.  
521 [arXiv:1709.09360](https://arxiv.org/abs/1709.09360) [cs.CL].
- 522 Zhang, Jiajun, Shujie Liu, Mu Li, Ming Zhou, and Chengqing Zong. 2014. Bilingually constrained  
523 phrase embeddings for machine translation. In *ACL*.

# Author Queries

Chapter 5

---

Query Refs.	Details Required	Author's response
	No queries.	

# Metadata of the chapter that will be visualized in SpringerLink

Book Title	Neural Representations of Natural Language	
Series Title		
Chapter Title	Conclusion	
Copyright Year	2019	
Copyright HolderName	Springer Nature Singapore Pte Ltd.	
Corresponding Author	Family Name Particle Given Name Prefix Suffix Role Division Organization Address Email	<b>White</b> <b>Lyndon</b> <b>Togneri</b> <b>Roberto</b> <b>Liu</b> <b>Wei</b> <b>Bennamoun</b> <b>Mohammed</b>
Author	School of Electrical, Electronic and Computer Engineering The University of Western Australia Perth, WA, Australia LYNDON.WHITE@RESEARCH.UWA.EDU.AU	
Author	School of Electrical, Electronic and Computer Engineering The University of Western Australia Perth, WA, Australia	
Author	School of Computer Science and Software Engineering The University of Western Australia Perth, WA, Australia	
Author		

Role  
Division School of Computer Science and Software Engineering  
Organization The University of Western Australia  
Address Perth, WA, Australia  
Email

---

Abstract

In this book we have introduced methods for finding representations of natural language. A special focus of this book is on neural networks and related technologies. Neural networks elegantly produce useful representations as by-products, when applied to NLP tasks. We have introduced recent advances in neural networks, in particular, various forms of recurrent neural networks, and have covered techniques for working with words, word senses, and larger structures such as phrases, sentences and documents.

---



# Conclusion

6

*The key to artificial intelligence has always been the representation. You and I are streaming data engines*

Jeff Hawkins 2012

## Some final reminders

This section has some final reminders, of points made earlier, and some final remarks not made elsewhere.

- 1 In this book we have introduced methods for finding representations of natural lan-  
2 guage. A special focus of this book is on neural networks and related technolo-  
3 gies. Neural networks elegantly produce useful representations as by-products, when  
4 applied to NLP tasks. We have introduced recent advances in neural networks, in  
5 particular, various forms of recurrent neural networks, and have covered techniques  
6 for working with words, word senses, and larger structures such as phrases, sentences  
7 and documents.

### Reminder: Gradient Checks

Back-propagation is a very easy algorithm to mess up, and it is hard to tell when you do mess it up, because it will often still work reasonably well. Gradient checks, either with finite-differencing, or with more sophisticated techniques (e.g. dual-numbers), is a quick and easy way to check the correctness of your implementations.

- 8 Chapter 2 introduced the idea of machine learning, which allows us to train a system using examples of the desired outputs for given inputs. The system can learn to



9 determine the outputs for inputs that were never seen during training. A key feature  
10 of modern machine learning is the decreased reliance on hand-engineered features  
11 to represent their inputs. The learnt embeddings and vector representations which  
12 are the main topic of this book, are just such automatically derived features.

### Hyperparameters matter

Neural networks have many hyper parameters, from hidden layer-size, to the regularisation penalty, to the number of training iterations. Choosing the right ones can significantly impact performance.

13 Chapter 3 showed how recurrent neural networks allowed working with inputs of  
14 varying lengths – such as natural language sentences. It discussed the various types of  
15 recurrent networks as characterised by their recurrent units, including both GRU and  
16 LSTM. As well as the common RNN structures, including: the encoder for producing  
17 a fixed-size output (such as a classification) from a varying sized input; the decoder  
18 for generating a varying sized output from a fixed size input; and the encoder-decoder  
19 for when the input and output both vary in size. These models are very practical for  
20 natural language processing applications.

### Preprocessing Matters

There is a large number of preprocessing tricks that can be employed when processing text. Options include removing stop-words, rare words, and punctuation, as well as removing or replacing numbers and dates. Different options for tokenization exist with regard to splitting up contractions and other factors. Often, it is good to convert all the text to lower-case. One can even lemmatize or stem every word occurrence. What is useful depends on the task.

21 Chapter 4 discussed how we can find and use representations of words. This is one  
22 of the most important ideas in machine learning for NLP. It begins with the core idea  
23 of how a word can be input into a neural network: the input embedding via a look-up  
24 table. The input embeddings, and the complementarity output embeddings obtained  
25 from the weights of the softmax layer, capture a representation of the words. The  
26 meaning of a word is determined by its usage, which is largely characterised by what  
27 words it co-occurs with. Word representation models include the well known skip-  
28 gram model which use these principles to derive high quality representations that can  
29 be reused as features in many tasks. The skip-gram model is based around predicting  
30 which words will be co-located. It is closely related to an iterative approximation  
31 to factorising the matrix of co-occurrence counts. This is similar to the approaches  
32 commonly used in older methods such as LSI and LDA. Word embedding models  
33 commonly use hierarchical softmax or negative sampling to speed-up the training  
34 and the evaluation.

**Reminder: There is more to clustering than K-means**

K-means is the most well-known clustering algorithm. But it is by no means the only one. K-means in particular is very vulnerable to getting stuck in local optima, compared to many other clustering algorithms. If one does use k-means, then make sure to run it multiple times and take the best result.

Chapter 5 extends the idea of representing words to representing senses. Most words have many meanings. It is thus impossible for a single representation to characterise the correct meaning in all contexts. Word senses can either be externally defined using a lexical resource, or discovered (induced) from the data. If the senses are externally defined, determining a good representation for them boils down to disambiguating a corpus to find the senses used, then creating a representation for them using the single sense word embedding methods. Inducing the senses from the corpus is more in-line with the general goal of not needing hand-engineered features, and there are many more methods for this.

The majority of word sense induction methods are either context-clustering-based, or co-location-prediction based. In both cases, the core idea is still that (like for word-embeddings) the surrounding words characterise what a particular word use means. In the *context clustering* methods, for each word the different contexts in which it occurs are represented and then clustered. Each cluster represents a word sense. In the *co-location prediction* approaches the word sense is treated as a hidden (latent) variable, which influencing the observable variables, i.e. the context words which it is co-occurring with. Classical probability methods can then be used, together with neural network methods, in order to uncover that latent variable that is the sense. These word sense induction methods give embeddings that can be used in much the same way as word embeddings.

**Machine learning beyond neural networks**

Though we have barely mentioned them in this book, there are plenty of other machine learning algorithms beyond just neural networks. Many of these have great advantages over neural networks, in terms of their performance with smaller amounts of data.

Some examples of some of the diversity of options (contrasted with neural networks) can be found at <https://white.ucc.asn.au/2017/12/18/7-Binary-Classifier-Libraries-in-Julia.html>.

It is often ideal to take word-embeddings (or other neural network derived representations) and use them as features, in a classifier such as a support vector machine (Cortes and Vapnik 1995), or a gradient boosting tree ensemble (Chen and Guestrin 2016).

Chapter 6 takes the idea of representing words to larger structures. There is a great diversity of methods to represent phrases, sentences, and documents. In three broad categories, we can consider weakly ordered models, sequential models, and structural

models. The *weakly ordered models* are the most diverse category, being a catch-all term for a variety of methods that do not directly consider word order, including the neural embedding extensions of bag of words and bag of n-grams. The *sequential methods* are largely the application of the RNNs from Chap. 3, but also include some methods specifically for obtaining a representation. Finally, the *structural methods* allow the inputs to change the structure of the network, allowing it to process parse trees in a natural way.

All things considered, there are a great number of techniques for using natural language with machine learning. They all revolve around the core idea of a representation. Beyond that they range from simple to complex.

The methods used need to be sufficiently powerful to accomplish the task, but beyond this excess capacity is both a waste of training time, and a waste of developer effort. Spending more time to implement one of the more complicated models may not result in a final system that even works as well as a simple baseline. The ideal system is suitably simple. However, the identification of what it means for an implementation to be suitably simple takes surprising amounts of research. In part this can perhaps be attributed to our over-expectation of complexity in language. While language can be complex, a lot of it is surprisingly simple.

The systems we have discussed have been small and self-contained. They are suitable for use as a component in a larger system. If one looks at the workings of a digital assistant (such as the ones found on a smart phone), one will find many machine learning based subsystems: for speech-recognition, for intent detection, and for accomplishing subtasks within those. A complicated system such as Zara (Siddique et al. 2017) contains over a dozen separate machine learning based subsystems.

### Zara, and defending herself from bad influences

One of the most interesting subsystems in Zara (Siddique et al. 2017), is the use of a module to detect abusive language, racism and sexism. It is a well-known problem that when allowing a learning system to learn from the open-web, such systems can pick-up bad habits (<https://www.theverge.com/2016/3/24/11297050>). The use of such an abuse detection module allows a system to be protected from such things.

This can be related to our earlier comments: with regards to the issues with WordNet's sexist definitions, and the issues that word embeddings can learn the biases representations from biases texts discussed by Bolukbasi et al. (2016) and Caliskan et al. (2017)

There is another approach, processing huge corpora of texts, with end-to-end deep learning systems. For example using decades of user support logs to train question answering systems. Even this is based on many of the same principles we have discussed in this book. Though, at this scale other options open up, like char-RNNs which function on a character scale rather than the word-scale that we have considered here. With enough data, all the linguistic modelling concerns can be learnt into the

neural network. The data and computational requirements place such things out of reach of many developers.

Most tasks do not have this scale of data. They may be small data, a few hundred examples. They may be largish data, with a few million examples. The more data you have, the more complex a model you can train, and thus the more difficult a problem you can solve. But good representations can help us when we do not have enough data. By creating word embeddings (Chap. 4), word sense embeddings (Chap. 5), and larger embeddings (Chap. 6), we can leverage other larger data sources. The embeddings created using systems that are trained on other data, to perform other tasks, capture information to accomplish those tasks. We can then take those information dense embeddings, and apply them to our own tasks.

In conclusion, this book has covered popular deep neural networks for language processing. The networks have been comprehensively reviewed and explained from a practical perspective. Language modelling, vector representations, and challenging tasks such as WSD, and sentence embeddings were investigated to illustrate the use of these networks. This book shares a lot of valuable advice obtained from first-hand experience in implementing these networks. We would treat this book as a solid introduction to one of the most exciting new areas of natural language processing and computational linguistics.

107

---

## References

- Bolukbasi, Tolga, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in neural information processing systems*, 4349–4357.
- Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356 (6334), 183–186. ISSN: 0036-8075, <https://doi.org/10.1126/science.aal4230>, <http://science.sciencemag.org/content/356/6334/183.full.pdf>.
- Chen, Tianqi and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 785–794. ACM.
- Cortes, Corinna and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning* 20 (3): 273–297. ISSN: 1573-0565, <https://doi.org/10.1007/BF00994018>.
- Siddique, Farhad Bin, Onno Kampman, Yang Yang, Anik Dey, and Pascale Fung. 2017. Zara returns: Improved personality induction and adaptation by an empathetic virtual agent. In *Proceedings of ACL 2017, system demonstrations*, 121–126.

# Author Queries

## Chapter 6

Query Refs.	Details Required	Author's response
	No queries.	

# Index

- A**
- 1 Acausal language modelling, 49, 100
  - 2 Activation function
    - 3 leaky-ReLU, 10, 19, 20
    - 4 ReLU, 4, 6, 9, 10
    - 5 ReLU6, 6, 9, 10, 20
    - 6 sigmoid, 4, 6–8, 60
    - 7 tanh, 6, 8
  - 8 Analogy, 53, 54, 95
  - 9 Autoencoder, 17, 19, 20, 98, 99, 104, 108,
  - 10 109
- B**
- 12 BabelNet, 75, 77, 88
  - 13 Bag of Words (BOW), 50, 68, 95, 96, 120
  - 14 Basic recurrent unit, 29, 47, 106
  - 15 Bayes' Theorem (Bayes' Law), 15, 17, 89
  - 16 Bidirectional recurrent neural network, 26,
  - 17 34
- C**
- 19 Canonical Correlation Analysis (CCA), 69
  - 20 Classifier, 7, 8, 14, 15, 87, 108, 119
  - 21 Clustering, 40, 73, 83–86, 119
  - 22 Co-location, 37, 57, 64, 83, 86, 119
  - 23 Compositional, 113
  - 24 Constituency parse tree, 100, 102, 105, 106,
  - 25 110
  - 26 Context, 26, 40, 43, 49–51, 53, 69, 74, 76,
  - 27 78, 81, 83–90, 94, 96, 119
  - 28 Continuous Bag of Words (CBOW), 50
  - 29 Corpus, 39, 40, 46, 53, 65, 77–79, 113, 119
- D**
- 31 Cosine distance (Cosine similarity), 79, 81
- E**
- 32 Decoder RNN, 27, 28, 34, 100, 109
  - 33 Deep recurrent neural network, 34
  - 34 Dependency parse tree, 100, 103, 105
- F**
- 36 Factorisation, 57, 83
  - 37 Features, 4, 16, 17, 37, 43, 53, 68, 74, 75,
  - 38 111, 118, 119
  - 39 Function approximation, 4, 76
- G**
- 44 Gated Recurrent Unit (GRU), 30
  - 45 GloVe, 50, 55–57
- H**
- 47 Hidden layer, 4–6, 9, 15, 16, 31, 43, 46, 50,
  - 48 53, 63, 65, 97, 118
  - 49 Hierarchical softmax, 15, 37, 46, 49, 50, 57–
  - 50 59, 61, 62, 64, 87
  - 51 Huffman tree, 57, 58
  - 52 Hyper-parameter, 4, 5, 46, 55, 67
- I**
- 54 Information retrieval, 37, 94, 95

- 56 Input embedding, 45, 53, 69, 82, 85, 87, 93,  
57 118
- 58 **L**
- 59 Language modelling, 37, 40, 49, 57, 62, 64,  
60 87, 100, 121
- 61 Long Short Term Memory (LSTM), 23, 30,  
62 106
- 63 Loss function, 11–13, 64–66, 82, 84, 109
- 64 **M**
- 65 Machine translation, 28, 97
- 66 Matched-Sequence RNN, 24–26, 28, 34, 98
- 67 Matrix vector model, 110, 111
- 68 Multilayer perceptron, 2
- 69 **N**
- 70 Nearest neighbour trees, 89
- 71 Negative sampling, 37, 50, 57, 64, 68, 87,  
72 107, 118
- 73 Neural Probabilistic Language Model  
(NPLM), 41, 46, 47
- 74 n-grams, 40, 46, 95, 97, 120
- 76 **O**
- 77 Output embedding, 16, 43–45, 53, 56, 82,  
78 86, 87, 118
- 79 **P**
- 80 Paragraph Vector Models (PV-DM, PV-  
81 DBOW), 96, 97
- 82 Parse tree, 100, 102–107, 120
- 83 Parsing, 100, 107–109
- 84 Probability, 8, 15–17, 26, 31, 39, 40, 51, 53,  
85 56, 58, 59, 61, 63–66, 77, 78, 86–  
86 89
- 87 **R**
- 88 Recurrent Neural Network (RNN), 23, 24,  
89 93, 97, 117
- 90 Recurrent Unit (RU), 24, 25, 29, 30, 33, 34,  
91 47, 97, 98, 106, 118
- 92 Recursive neural network, 102
- 93 **S**
- 94 Skip-gram, 37, 43, 50–53, 55–57, 62, 65, 68,  
95 79, 81–86, 95–97, 118
- 96 Skip-thought, 101
- 97 Softmax, 6–8, 15, 16, 27, 37, 43, 46, 49, 50,  
98 52, 57–59, 61, 63–65, 87, 88, 93,  
99 108, 118
- 100 Structured network (Structured model), 104,  
101 105, 107, 108, 110
- 102 Sum of Word Embeddings (SOWE), 93, 95,  
103 110, 114
- 104 Supervised learning, 109
- 105 **T**
- 106 Training, 2, 4, 6, 10–12, 15, 23, 26, 28, 30,  
107 40, 41, 43, 46, 52, 53, 55, 57, 58,  
108 63–67, 74, 76, 77, 79, 81, 85–87,  
109 96, 99, 104, 109, 118
- 110 Trigrams, 40, 95
- 111 **U**
- 112 Unsupervised learning, 5, 38
- 113 **V**
- 114 Variational Autoencoder (VAE) , 98
- 115 **W**
- 116 Window, 46, 50, 53, 55, 84, 96, 97
- 117 WordNet, 74–76, 78, 79, 88, 120
- 118 Word Sense Disambiguation (WSD), 75, 77,  
119 78, 82, 94
- 120 Word Sense Induction (WSI), 73, 77, 83, 119
- 121 Word sense representation, 73, 79, 90
- 122 Word2Vec, 37, 51
- 123 **Z**
- 124 Zipf's law, 65, 77



# Author Queries

Chapter

---

Query Refs.	Details Required	Author's response
	No queries.	

# MARKED PROOF

## Please correct and return this set

Please use the proof correction marks shown below for all alterations and corrections. If you wish to return your proof by fax you should ensure that all amendments are written clearly in dark ink and are made well within the page margins.

<b>Instruction to printer</b>	<b>Textual mark</b>	<b>Marginal mark</b>
Leave unchanged	... under matter to remain	✓
Insert in text the matter indicated in the margin	↪	New matter followed by ↪ or ↪ <sup>x2</sup>
Delete	/ through single character, rule or underline ———— through all characters to be deleted	⊖ or ⊖ <sup>x2</sup>
Substitute character or substitute part of one or more word(s)	/ through letter or ———— through characters	new character / or new characters /
Change to italics	— under matter to be changed	←
Change to capitals	≡ under matter to be changed	≡
Change to small capitals	= under matter to be changed	=
Change to bold type	~ under matter to be changed	~
Change to bold italic	~~ under matter to be changed	~~
Change to lower case	Encircle matter to be changed	≠
Change italic to upright type	(As above)	+
Change bold to non-bold type	(As above)	—
Insert 'superior' character	/ through character or ↪ where required	Y or X under character e.g. ³ or ³
Insert 'inferior' character	(As above)	↪ over character e.g. ²
Insert full stop	(As above)	◎
Insert comma	(As above)	,
Insert single quotation marks	(As above)	‘ or ’ and/or ‘ or ’
Insert double quotation marks	(As above)	“ or ” and/or “ or ”
Insert hyphen	(As above)	—
Start new paragraph	—	—
No new paragraph	≈	≈
Transpose	□	□
Close up	linking ( ) characters	( )
Insert or substitute space between characters or words	/ through character or ↪ where required	Y
Reduce space between characters or words	between characters or words affected	T