

Contents

	Preface	xvii
	Acknowledgments	xxi
1	Introduction	1
1.1	The Challenges of Natural Language Processing	1
1.2	Neural Networks and Deep Learning	2
1.3	Deep Learning in NLP	3
1.3.1	Success Stories	4
1.4	Coverage and Organization	6
1.5	What's not Covered	8
1.6	A Note on Terminology	9
1.7	Mathematical Notation	9
	 PART I Supervised Classification and Feed-forward Neural Networks	 11
2	Learning Basics and Linear Models	13
2.1	Supervised Learning and Parameterized Functions	13
2.2	Train, Test, and Validation Sets	14
2.3	Linear Models	16
2.3.1	Binary Classification	16
2.3.2	Log-linear Binary Classification	20
2.3.3	Multi-class Classification	21
2.4	Representations	22
2.5	One-Hot and Dense Vector Representations	23
2.6	Log-linear Multi-class Classification	24
2.7	Training as Optimization	25
2.7.1	Loss Functions	26
2.7.2	Regularization	29

2.8	Gradient-based Optimization	30
2.8.1	Stochastic Gradient Descent	31
2.8.2	Worked-out Example	33
2.8.3	Beyond SGD	35
3	From Linear Models to Multi-layer Perceptrons	37
3.1	Limitations of Linear Models: The XOR Problem	37
3.2	Nonlinear Input Transformations	38
3.3	Kernel Methods	38
3.4	Trainable Mapping Functions	39
4	Feed-forward Neural Networks	41
4.1	A Brain-inspired Metaphor	41
4.2	In Mathematical Notation	43
4.3	Representation Power	44
4.4	Common Nonlinearities	45
4.5	Loss Functions	46
4.6	Regularization and Dropout	47
4.7	Similarity and Distance Layers	48
4.8	Embedding Layers	49
5	Neural Network Training	51
5.1	The Computation Graph Abstraction	51
5.1.1	Forward Computation	53
5.1.2	Backward Computation (Derivatives, Backprop)	54
5.1.3	Software	55
5.1.4	Implementation Recipe	57
5.1.5	Network Composition	58
5.2	Practicalities	58
5.2.1	Choice of Optimization Algorithm	59
5.2.2	Initialization	59
5.2.3	Restarts and Ensembles	59
5.2.4	Vanishing and Exploding Gradients	60
5.2.5	Saturation and Dead Neurons	60
5.2.6	Shuffling	61
5.2.7	Learning Rate	61
5.2.8	Minibatches	61

PART II Working with Natural Language Data 63

6	Features for Textual Data	65
6.1	Typology of NLP Classification Problems	65
6.2	Features for NLP Problems	67
6.2.1	Directly Observable Properties	67
6.2.2	Inferred Linguistic Properties	70
6.2.3	Core Features vs. Combination Features	74
6.2.4	Ngram Features	75
6.2.5	Distributional Features	76
7	Case Studies of NLP Features	77
7.1	Document Classification: Language Identification	77
7.2	Document Classification: Topic Classification	77
7.3	Document Classification: Authorship Attribution	78
7.4	Word-in-context: Part of Speech Tagging	79
7.5	Word-in-context: Named Entity Recognition	81
7.6	Word in Context, Linguistic Features: Preposition Sense Disambiguation	82
7.7	Relation Between Words in Context: Arc-Factored Parsing	85
8	From Textual Features to Inputs	89
8.1	Encoding Categorical Features	89
8.1.1	One-hot Encodings	89
8.1.2	Dense Encodings (Feature Embeddings)	90
8.1.3	Dense Vectors vs. One-hot Representations	90
8.2	Combining Dense Vectors	92
8.2.1	Window-based Features	93
8.2.2	Variable Number of Features: Continuous Bag of Words	93
8.3	Relation Between One-hot and Dense Vectors	94
8.4	Odds and Ends	95
8.4.1	Distance and Position Features	95
8.4.2	Padding, Unknown Words, and Word Dropout	96
8.4.3	Feature Combinations	98
8.4.4	Vector Sharing	98
8.4.5	Dimensionality	99
8.4.6	Embeddings Vocabulary	99
8.4.7	Network's Output	99

8.5	Example: Part-of-Speech Tagging	100
8.6	Example: Arc-factored Parsing	101
9	Language Modeling	105
9.1	The Language Modeling Task	105
9.2	Evaluating Language Models: Perplexity	106
9.3	Traditional Approaches to Language Modeling	107
9.3.1	Further Reading	108
9.3.2	Limitations of Traditional Language Models	108
9.4	Neural Language Models	109
9.5	Using Language Models for Generation	112
9.6	Byproduct: Word Representations	113
10	Pre-trained Word Representations	115
10.1	Random Initialization	115
10.2	Supervised Task-specific Pre-training	115
10.3	Unsupervised Pre-training	116
10.3.1	Using Pre-trained Embeddings	117
10.4	Word Embedding Algorithms	117
10.4.1	Distributional Hypothesis and Word Representations	118
10.4.2	From Neural Language Models to Distributed Representations	122
10.4.3	Connecting the Worlds	125
10.4.4	Other Algorithms	126
10.5	The Choice of Contexts	127
10.5.1	Window Approach	127
10.5.2	Sentences, Paragraphs, or Documents	129
10.5.3	Syntactic Window	129
10.5.4	Multilingual	130
10.5.5	Character-based and Sub-word Representations	131
10.6	Dealing with Multi-word Units and Word Inflections	132
10.7	Limitations of Distributional Methods	133
11	Using Word Embeddings	135
11.1	Obtaining Word Vectors	135
11.2	Word Similarity	135
11.3	Word Clustering	136

11.4	Finding Similar Words	136
11.4.1	Similarity to a Group of Words	137
11.5	Odd-one Out	137
11.6	Short Document Similarity	137
11.7	Word Analogies	138
11.8	Retrofitting and Projections	139
11.9	Practicalities and Pitfalls	140
12	Case Study: A Feed-forward Architecture for Sentence Meaning Inference ..	141
12.1	Natural Language Inference and the SNLI Dataset	141
12.2	A Textual Similarity Network	142
 PART III Specialized Architectures		147
13	Ngram Detectors: Convolutional Neural Networks	151
13.1	Basic Convolution + Pooling	152
13.1.1	1D Convolutions Over Text	153
13.1.2	Vector Pooling	155
13.1.3	Variations	158
13.2	Alternative: Feature Hashing	158
13.3	Hierarchical Convolutions	159
14	Recurrent Neural Networks: Modeling Sequences and Stacks	163
14.1	The RNN Abstraction	164
14.2	RNN Training	166
14.3	Common RNN Usage-patterns	167
14.3.1	Acceptor	167
14.3.2	Encoder	167
14.3.3	Transducer	168
14.4	Bidirectional RNNs (biRNN)	169
14.5	Multi-layer (stacked) RNNs	171
14.6	RNNs for Representing Stacks	172
14.7	A Note on Reading the Literature	174

15	Concrete Recurrent Neural Network Architectures	177
15.1	CBOW as an RNN	177
15.2	Simple RNN	177
15.3	Gated Architectures	178
15.3.1	LSTM	179
15.3.2	GRU	181
15.4	Other Variants	182
15.5	Dropout in RNNs	183
16	Modeling with Recurrent Networks	185
16.1	Acceptors	185
16.1.1	Sentiment Classification	185
16.1.2	Subject-verb Agreement Grammaticality Detection	187
16.2	RNNs as Feature Extractors	189
16.2.1	Part-of-speech Tagging	189
16.2.2	RNN-CNN Document Classification	191
16.2.3	Arc-factored Dependency Parsing	192
17	Conditioned Generation	195
17.1	RNN Generators	195
17.1.1	Training Generators	196
17.2	Conditioned Generation (Encoder-Decoder)	196
17.2.1	Sequence to Sequence Models	198
17.2.2	Applications	200
17.2.3	Other Conditioning Contexts	202
17.3	Unsupervised Sentence Similarity	203
17.4	Conditioned Generation with Attention	204
17.4.1	Computational Complexity	208
17.4.2	Interpretability	208
17.5	Attention-based Models in NLP	208
17.5.1	Machine Translation	209
17.5.2	Morphological Inflection	210
17.5.3	Syntactic Parsing	211

PART IV	Additional Topics	213
18	Modeling Trees with Recursive Neural Networks	215
18.1	Formal Definition	215
18.2	Extensions and Variations	218
18.3	Training Recursive Neural Networks	219
18.4	A Simple Alternative—Linearized Trees	219
18.5	Outlook	220
19	Structured Output Prediction	221
19.1	Search-based Structured Prediction	221
19.1.1	Structured Prediction with Linear Models	221
19.1.2	Nonlinear Structured Prediction	222
19.1.3	Probabilistic Objective (CRF)	224
19.1.4	Approximate Search	224
19.1.5	Reranking	225
19.1.6	See Also	225
19.2	Greedy Structured Prediction	226
19.3	Conditional Generation as Structured Output Prediction	227
19.4	Examples	228
19.4.1	Search-based Structured Prediction: First-order Dependency Parsing	228
19.4.2	Neural-CRF for Named Entity Recognition	229
19.4.3	Approximate NER-CRF With Beam-Search	232
20	Cascaded, Multi-task and Semi-supervised Learning	235
20.1	Model Cascading	235
20.2	Multi-task Learning	238
20.2.1	Training in a Multi-task Setup	241
20.2.2	Selective Sharing	242
20.2.3	Word-embeddings Pre-training as Multi-task Learning	243
20.2.4	Multi-task Learning in Conditioned Generation	243
20.2.5	Multi-task Learning as Regularization	243
20.2.6	Caveats	244
20.3	Semi-supervised Learning	244
20.4	Examples	245
20.4.1	Gaze-prediction and Sentence Compression	245
20.4.2	Arc Labeling and Syntactic Parsing	246

20.4.3	Preposition Sense Disambiguation and Preposition Translation Prediction	247
20.4.4	Conditioned Generation: Multilingual Machine Translation, Parsing, and Image Captioning	249
20.5	Outlook	250
21	Conclusion	251
21.1	What Have We Seen?	251
21.2	The Challenges Ahead	251
	Bibliography	253
	Author's Biography	287