

Contents

1.	Introduction	1
1.1	What Is Chinese NLP?.....	1
1.2	About This Book.....	5
2.	Words in Chinese	7
2.1	Introduction.....	7
2.2	Characters, Morphemes, and Words	7
2.2.1	Characters (字).....	7
2.2.2	Morphemes (词素).....	8
2.2.3	Words (词).....	9
2.3	Word Formation in Chinese.....	11
2.3.1	Disyllabic Compounds	12
2.3.2	Trisyllabic Compounds.....	18
2.3.3	Quadrasyllabic Compounds.....	20
2.3.4	Other Morphological Processes in Chinese.....	24
2.3.5	Ionization (离合).....	25
2.4	Word Identification and Segmentation	25
2.5	Summary	25
3.	Challenges in Chinese Morphological Processing	27
3.1	Introduction.....	27
3.2	Chinese Characters.....	27
3.2.1	Large Number of Characters.....	27
3.2.2	Simplified and Traditional Characters	28
3.2.3	Variant Characters	28
3.2.4	Dialect Characters and Dialectal Use of Standard Characters.....	28
3.2.5	Multiple Character Encoding Standards.....	30
3.3	Textual Conventions.....	32
3.3.1	Printing Format.....	32
3.3.2	Punctuation Practice.....	33

3.4	Linguistic Characteristics	33
3.4.1	Few Formal Morphological Markings	33
3.4.2	Parts of Speech	34
3.4.3	Homonyms and Homographs	36
3.4.4	Ambiguity.....	37
3.4.5	OOV Words 未登录词	39
3.4.6	Regional Variation	40
3.4.7	Stylistic Variation	40
3.5	Summary	40
4.	Chinese Word Segmentation	41
4.1	Introduction.....	41
4.2	Two Main Challenges	41
4.3	Algorithms.....	43
4.3.1	Character-Based Approach	43
4.3.2	Word-Based Approach.....	47
4.4	Word Segmentation Ambiguity	54
4.4.1	Ambiguity Definition	54
4.4.2	Disambiguation Algorithms.....	55
4.5	Benchmarks	57
4.5.1	Standards.....	57
4.5.2	Bakeoff Evaluation	58
4.6	Free Tools	59
4.6.1	Chinese Lexical Analysis System	60
4.6.2	MSRSeg System.....	60
4.7	Summary	60
5.	Unknown Word Identification	61
5.1	Introduction.....	61
5.2	Unknown Word Detection and Recognition.....	63
5.3	Chinese Person Name Identification	66
5.4	Chinese Organization Name Identification.....	68
5.5	Chinese Place Name Recognition	70
5.6	Summary	71
6.	Word Meaning.....	73
6.1	Basic Meaning, Concept, and Relation	73

6.2	Frame, Collocation, and Verb Valency.....	74
6.3	Chinese Dictionaries/Lexicons.....	75
6.3.1	CILIN (《同义词词林》).....	76
6.4	WordNets.....	78
6.4.1	HowNet.....	78
6.4.2	Chinese Concept Dictionary.....	81
6.5	Summary.....	94
7.	Chinese Collocations.....	95
7.1	Concept of Collocation.....	95
7.1.1	Definitions.....	95
7.1.2	Collocations in Chinese.....	97
7.2	Qualitative Properties.....	98
7.3	Quantitative Features.....	100
7.4	Categorization.....	101
7.5	Linguistic Resources.....	103
7.5.1	Collocation Dictionary of Modern Chinese Lexical Words.....	103
7.5.2	Dictionary of Modern Chinese Collocations.....	105
7.5.3	Chinese Collocation Bank.....	105
7.6	Applications.....	107
7.7	Summary.....	108
8.	Automatic Chinese Collocation Extraction.....	109
8.1	Introduction.....	109
8.2	Window-Based Statistical Approach.....	109
8.2.1	Co-occurrence Frequency Features.....	110
8.2.2	Co-occurrence Distribution Features.....	118
8.2.3	Context-Based Features.....	120
8.2.4	Window Size.....	122
8.3	Syntax-Based Approach.....	122
8.3.1	POS Pattern Filter.....	123
8.3.2	Dependency Triple-Based Extraction.....	124
8.3.3	Using Chunk Information.....	125
8.4	Semantic-Based Approach.....	126
8.4.1	Synonym Substitution Test.....	126
8.4.2	Translation Testing.....	127
8.5	Categorization-Based Approach.....	127

8.6	Benchmarks.....	129
8.6.1	Dictionary of Modern Chinese Collocations (Electronic Version).....	129
8.6.2	Collocations for 100 Chinese Headwords.....	130
8.7	Summary.....	130
Appendix.....		131
References.....		135
Author Biographies.....		147