

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
28		
29		
30		
31		
32		
33		
34		
35		
36		
37		
38		
39		
40		
41		
42		
43		
44		
45		
46		
47		
48		
49		
50		
51		
52		
53		
54		
55		
56		
57		
58		
59		
60		
61		
62		
63		
64		
65		
66		
67		
68		
69		
70		
71		
72		
73		
74		
75		
76		
77		
78		
79		
80		
81		
82		
83		
84		
85		
86		
87		
88		
89		
90		
91		
92		
93		
94		
95		
96		
97		
98		
99		
100		
101		
102		
103		
104		
105		
106		
107		
108		
109		
110		
111		
112		
113		
114		
115		
116		
117		
118		
119		
120		
121		
122		
123		
124		
125		
126		
127		
128		
129		
130		
131		
132		
133		
134		
135		
136		
137		
138		
139		
140		
141		
142		
143		
144		
145		
146		
147		
148		
149		
150		
151		
152		
153		
154		
155		
156		
157		
158		
159		
160		
161		
162		
163		
164		
165		
166		
167		
168		
169		
170		
171		
172		
173		
174		
175		
176		
177		
178		
179		
180		
181		
182		
183		
184		
185		
186		
187		
188		
189		
190		
191		
192		
193		
194		
195		
196		
197		
198		
199		
200		
201		
202		
203		
204		
205		
206		
207		
208		
209		
210		
211		
212		
213		
214		
215		
216		
217		
218		
219		
220		
221		
222		
223		
224		
225		
226		
227		
228		
229		
230		
231		
232		
233		
234		
235		
236		
237		
238		
239		
240		
241		
242		
243		
244		
245		
246		
247		
248		
249		
250		
251		
252		
253		
254		
255		
256		
257		
258		
259		
260		
261		
262		
263		
264		
265		
266		
267		
268		
269		
270		
271		
272		
273		
274		
275		
276		
277		
278		
279		
280		
281		
282		
283		
284		
285		
286		
287		
288		
289		
290		
291		
292		
293		
294		
295		
296		
297		
298		
299		
300		
301		
302		
303		
304		
305		
306		
307		
308		
309		
310		
311		
312		
313		
314		
315		
316		
317		
318		
319		
320		
321		
322		
323		
324		
325		
326		
327		
328		
329		
330		
331		
332		
333		
334		
335		
336		
337		
338		
339		
340		
341		
342		
343		
344		
345		
346		
347		
348		
349		
350		
351		
352		
353		
354		
355		
356		
357		
358		
359		
360		
361		
362		
363		
364		
365		
366		
367		
368		
369		
370		
371		
372		
373		
374		
375		
376		
377		
378		
379		
380		
381		
382		
383		
384		
385		
386		
387		
388		
389		
390		
391		
392		
393		
394		
395		
396		
397		
398		
399		
400		
401		
402		
403		
404		
405		
406		
407		
408		
409		
410		
411		
412		
413		
414		
415		
416		
417		
418		
419		
420		
421		
422		
423		
424		
425		
426		
427		
428		
429		
430		
431		
432		
433		
434		
435		
436		
437		
438		
439		
440		
441		
442		
443		
444		
445		
446		
447		
448		
449		
450		
451		
452		
453		
454		
455		
456		
457		
458		
459		
460		
461		
462		
463		
464		
465		
466		
467		
468		
469		
470		
471		
472		
473		
474		
475		
476		
477		
478		
479		
480		
481		
482		
483		
484		
485		
486		
487		
488		
489		
490		
491		
492		
493		
494		
495		
496		
497		
498		
499		
500		
501		
502		
503		
504		
505		
506		
507		
508		
509		
510		
511		
512		
513		
514		
515		
516		
517		
518		
519		
520		
521		
522		
523		
524		
525		
526		
527		
528		
529		
530		
531		
532		
533		
534		
535		
536		
537		
538		
539		
540		
541		
542		
543		
544		
545		
546		
547		
548		
549		
550		
551		
552		
553		
554		
555		
556		
557		
558		
559		
560		
561		
562		
563		
564		
565		
566		
567		
568		
569		
570		
571		
572		
573		
574		
575		
576		
577		
578		
579		
580		
581		
582		
583		
584		
585		
586		
587		
588		
589		
590		
591		
592		
593		
594		
595		
596		
597		
598		
599		
600		
601		
602		
603		
604		
605		
606		
607		