

Contents

摘要.....	iii
Abstract.....	v
Acknowledgements	vii
List of Figures.....	xii
List of Tables	xiv
Chapter 1: Introduction	1
1.1 Motivation.....	1
1.2 Our Approach to the Acquisition of NE Pairs	3
1.3 A Preview of Subsequent Chapters.....	4
Chapter 2: Related Work.....	7
2.1 NE Identification.....	8
2.2 Machine Transliteration.....	10
2.3 Bilingual Lexicon Construction.....	13
2.4 NE Alignment/Translation.....	14
Chapter 3: Bilingual NE Alignment.....	19
3.1 Problem Statement.....	19

3.2 Outline of the Proposed Approach.....	21
3.3 Statistical Phrase Translation Model (SPTM)	22
3.3.1 An Illustrative Example	24
3.3.2 Estimation of LTP and PAP.....	26
3.4 Abbreviation Handling (AH)	28
3.5 Chinese Person Name Recognition (CPNR).....	30
3.6 Acronym Expansion (AE)	34
Chapter 4: Machine Transliteration.....	39
4.1 Overview of the Noisy Channel Model	39
4.2 Formal Description: Transliteration Model (TM)	42
4.3 Estimation of Model Parameters.....	44
4.4 Alignment of transliteration pairs in parallel corpora.....	47
4.4.1 Overall Process	47
4.4.2 Linguistic Processing	51
4.5 Work Flow of Integrating Linguistic and Statistical Information.....	54
Chapter 5: Integrated Approach to NE Alignment.....	56
5.1 Framework of the Proposed Approach	56
5.2 Process of Aligning Bilingual NEs in Parallel Corpora	58

Chapter 6: Experiments on NE Alignment	65
6.1 Experimental Setup	65
6.2 Experimental Results and Discussion	67
Chapter 7: Experiments on Transliteration Alignment.....	84
7.1 Experimental Setup	84
7.2 TUs for English and Chinese	86
7.3 Evaluation Metric	91
7.4 Experimental Results and Discussion	92
Chapter 8: Conclusions and Future Work.....	98
8.1 Contributions.....	98
8.2 Future Work.....	100
Bibliography	102
Publications	112

List of Figures

Figure 3.1 The outline of the NE alignment process in parallel corpora.	21
Figure 3.2 Translation candidates with associated scores.	25
Figure 3.3 Transformation of an NE pair with approximate matching operations.	30
Figure 3.4 Verification process for the Chinese given name “靜茹.”	33
Figure 3.5 Alignment of the NE pair (Gypsy Chang, 張靜茹).	34
Figure 4.1 The noisy channel model in machine transliteration.	40
Figure 4.2 TU alignment between English and Chinese romanized character sequences.	41
Figure 4.3 The overall process for extracting name and transliteration pairs from parallel corpora.	48
Figure 4.4 The alignments of the TU matching pairs via the Viterbi path.	50
Figure 4.5 The Viterbi alignment path.	50
Figure 5.1 The framework for aligning bilingual NEs in parallel corpora.	58
Figure 5.2 The process of aligning bilingual NEs in parallel corpora.	59

Figure 5.3 Alignment of the NE pair (Juilliard School, 茱麗亞音樂學院).	63
Figure 6.1 Plot of the statistics in Table 6.8.	74
Figure 7.1 TU alignment of the name pair (Jacqueline, Chiehkuailin “傑桂琳”).	89
Figure 7.2 TU alignment of “Beaufort” and corresponding transliterations.	90



List of Tables

Table 3.2 A PAP table for “Ichthyosis Concern Association.”	26
Table 4.1 A portion of the list for translation.	52
Table 5.1 Examples of NE pairs in aligned sentences.	60
Table 5.2 Sets of Chinese NE candidates.	62
Table 6.1 Statistics of the <i>Sinorama</i> corpus.....	66
Table 6.2 Occurrence statistics for bilingual NE pairs in the <i>Sinorama</i> test set.	66
Table 6.3 Performance in bilingual NE alignment with the <i>Sinorama</i> test set.	67
Table 6.4 Average lengths of the NE types for the answer set in the <i>Sinorama</i> test set.	69
Table 6.5 Examples of possible Chinese NEs extracted by the proposed approach.	71
Table 6.6 Occurrence statistics for bilingual NE pairs in the <i>HKNPT</i> test set.	72
Table 6.7 Average lengths of the NE types for the answer set in the <i>HKNPT</i> test set.	72
Table 6.8 Performance in bilingual NE alignment with the <i>HKNPT</i> test set.....	73
Table 6.9 Examples of alignment errors made using the proposed approach.	77

Table 6.9 Examples of alignment errors made using the proposed approach. (Cont.)	78
Table 6.10 Performance for each language-specific knowledge source in the two corpora.....	79
Table 6.11 Detailed statistics on the numbers of translations and transliterations in the two corpora.....	79
Table 7.1 Some samples from the training set <i>T0</i>	85
Table 7.2 Some bilingual examples from the testing set <i>P1</i>	87
Table 7.3 Some high frequency English TUs.	88
Table 7.4 Some high frequency Chinese TUs.....	88
Table 7.5 English-Chinese TU-mapping probabilities.....	88
Table 7.6 Examples for each match type.....	92
Table 7.7 The experimental results of transliterated word extraction.....	93
Table 7.8 Some examples of Chinese transliterations, correctly extracted by the TM model, from <i>P1</i>	95
Table 7.9 Some examples of possible Chinese transliterations extracted by the proposed approaches.	96
Table 7.10 The average rates of transliterated word extraction for overall corpora.	97