Cross Sentence Alignment based on Singular Value Decomposition

by

Anna Ho

Master of Science in Software Engineering

2008



Faculty of Science and Technology
University of Macau

TABLE OF CONTENTS

List of Figures iv
List of Tables v
List of Abbreviations vi
Acknowledgements vii
Chapter 1: Introduction 1
1.1 Definition of Alignment 2
1.2 Literature review of Sentence Alignment Algorithms 3
1.2.1 Lexical Based Approach 4
1.2.2 Statistical Based Approach 7
1.2.3 Hybrid Based 11
1.3 Application Areas 14
1.3.1 Lexicography and Terminology 15
1.3.2 Machine Translation 15
1.3.3 Cross Language Information Extraction 16
1.4 Problem Statements 16
1.4.1 Linguistic Differentiation of Portuguese and Chinese 17
1.4.2 Limited Resources in Portuguese and Chinese Parallel Corpus 17
1.4.3 Deviations in Lexical Presentation 18
1.5 Objectives 19
1.6 Thesis Layout 20
Chapter 2: Alignment of Bilingual Corpus 23
2.1 Types of Bilingual Corpus 23
2.1.1 Parallel Corpus 23
2.1.2 Comparable Corpus 24
2.1.3Quasi-comparable Corpus 25
2.2 Levels of Alignment 25
2.2.1 Paragraph Level 25
2.2.2 Sentence Level 26

2.2.3 Phrase Level 26
2.2.4 Word Level 27
2.3 Chapter Conclusion 27
Chapter 3: Lexical Based with Non-match Penalty Approach 28
3.1 Advantages of Employing Lexicon Information 28
3.2 Morphological Analysis 29
3.3 Paired Lexicon Identification Module 29
3.3.1 Lexicon Matching Cases 30
3.3.2 Mathematical Model 31
3.4 Limitations 32
3.4.1 Limited Bilingual Resources 33
3.4.2 Sub-Phrase Problem 33
3.5 Non-Matched Penalty Module 34
3.6 Overview of the lexical based Alignment Framework 36
3.7 Portuguese-Chinese Alignment Matrix 37
3.8 Chapter Conclusion 38
Chapter 4: SVD Approach for Cross Alignment Problem 39
4.1 Highlighted Problems 39
4.1.1 Latent Semantic of Sentences 40
4.1.2 Disturbance of Non-Informative Lexicons 40
4.1.3 Confusion in Handling Cross Sentence Alignment with Similar Meanings 41
4.1.4 Overview of the Extended Alignment Framework 42
4.2 Elimination of Stop Words 44
4.3 Application of Singular Value Decomposition Technique 44
4.3.1 Construction of Statistical Relationship Matrix 46
4.3.2 Singular Value Decomposition 49
4.3.3 Similarity Measurement 51
4.3.4 Advantage of Dimensional Reduction in SVD 52
4.4 Chapter Conclusion 54

Chapter 5: SVD Based Approach with Enhanced Statistical Relationship Score 55
5.1 Inadequate information from Statistical Relationship Score 56
5.1.1 Sequential Order of Chinese Characters 56
5.1.2 Importance and Difficulties of Employing Special Alignment Clues 57
5.2 bilingual Documents Segmentation 58
5.3 scoring Features 60
5.3.1 Lexical Matching based on N-Gram 61
5.3.2 Length Ratio Measurement 63
5.3.3 Punctuation Matching 64
5.3.4 Digit Groups Matching 65
5.4 Assignment of Feature Weights 66
5.5 Chapter Conlcusion 66
Chapter 6: Evaluation Results 68
6.1 Evaluation Matrics 68
6.1.1 Notion of Recall 68
6.1.2 Notion of Precision 69
6.2 Testing Environment and Resources Used 69
6.2.1 Program Interface 69
6.2.2 Choice of Reference Corpus 71
6.3 Evaluation results With Recall 73
6.4 Evaluation Results with Precision 74
6.4.1 Average Precision 74
6.4.2 Precision of 70% 76
6.5 Chapter Conlcusion 77
Chapter 7: Conclusion 79
7.1 Research Contributions 80
7.2 Future Directions 81
References 82