

457
e32
JFON

**Research and Implementation of Chinese Segmentation Algorithms
and its Application to Chinese-Portuguese Machine Translation
System**

by

Leong Ka Seng

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	vii
Chapter 1: Introduction	1
1.1 Chinese Word Segmentation in Chinese Information Processing	2
1.2 Background of Chinese Word Segmentation.....	3
1.3 Difficulties in Chinese Word Segmentation	4
1.4 Problems in Chinese Word Segmentation Algorithm.....	6
1.4.1 Segmentation Ambiguities.....	6
1.4.2 Unknown Words	9
1.5 Approaches to Chinese Word Segmentation	11
1.5.1 Rule-based Approaches	12
1.5.2 Corpus-based Approaches	16
1.5.2.1 Statistics-based Methods.....	17
1.5.2.2 Machine Learning based Methods	22
1.6 Evaluation Criteria for Word Segmentation System	24
1.6.1 Accuracy	25
1.6.2 Speed.....	26
1.6.3 Generality and Applicability	27
1.7 Research Background and Direction	27
1.8 Thesis Layout and Brief Overview of Chapters	30
Chapter 2: CSAT- Integrated Chinese Segmentation and Tagging Module.....	33
2.1 Introduction.....	33
2.2 Rough Word Segmentation Model based on N-Shortest-Paths Model	35
2.3 Probabilistic Tagging Model based on HMM	37
2.4 Experiments	39
2.4.1 Evaluation on the Overall Performance of CSAT	39
2.4.1.1 Experiment One	40

2.4.1.2	Experiment Two.....	41
2.4.1.3	Discussion of Results.....	42
2.4.2	Evaluation of the Rough Segmentation on the Corpora of The First International Chinese Word Segmentation Bakeoff.....	43
2.5	Chapter Conclusion.....	46
Chapter 3: MESEG - Maximum Entropy based Word Segmenter		48
3.1	Introduction.....	48
3.2	Maximum Entropy Modeling	51
3.3	Word Segmentation as a Tagging Problem	54
3.4	Word Segmentation Features	57
3.5	Experiments	61
3.5.1	Without Post Processing to the Word Segmentation Result.....	61
3.5.2	With Post Processing to The Word Segmentation Result.....	63
3.6	Chapter Conclusion.....	65
Chapter 4: Integration of Named Entity Information for Word Segmentation.....		66
4.1	Introduction.....	66
4.2	Named Entity Identification and Related Work.....	67
4.3	UNEI as Tagging Problem.....	70
4.4	Features Formulation	73
4.4.1	Word Segmentation Features	74
4.4.2	UNEI Features	76
4.5	Experiments and Discussion	77
4.5.1	Experiment One	77
4.5.2	Experiment Two.....	78
4.5.3	Experiment Three.....	79
4.6	Chapter Conclusion.....	80
Chapter 5: Maximum Entropy Modeling Package		82
5.1	Introduction.....	82
5.2	Construction of the Maximum Entropy Modeling Package	83
5.2.1	Data Collection	84
5.2.2	Training Process.....	89

- 5.2.2.1 Training of Feature Weights 90
 - 5.2.2.2 Calculation of Probability Values 91
 - 5.2.2.3 Runtime Data Structures for Training Data 92
 - 5.2.2.4 IIS Implementation Issues.....95
- 5.3 Performance of the Maximum Entropy Modeling Package 100
 - 5.3.1 Experiment One 100
 - 5.3.2 Experiment Two.....102
- 5.4 Chapter Conclusion.....103
- Chapter 6: Application and Future Work.....105
 - 6.1 Construction of Open Track Word Segmenter 105
 - 6.1.1 External Knowledge for Unknown Named Entity 106
 - 6.1.1.1 Person Name 107
 - 6.1.1.2 Location Name.....109
 - 6.1.1.3 Organization Name 109
 - 6.1.2 External Knowledge for MDW.....111
 - 6.1.3 External Knowledge for FT 112
 - 6.2 Application to Chinese-Portuguese machine Translation System 113
 - 6.2.1 Word Segmentation 114
 - 6.2.2 POS Tagging 115
 - 6.2.3 Parsing Constraint-based Synchronous Grammar 115
 - 6.2.4 Target Translation Generation 116
 - 6.3 Application of Maximum Entropy Modeling Package to Other Areas 118
 - 6.4 Chapter Conclusion.....119
- Chapter 7: Conclusion.....120
- References.....124