

**Unsupervised Word Sense Disambiguation Using Non-Aligned  
Bilingual Corpus in Application to Portuguese-Chinese Machine  
Translation**

by

**Francisco de Oliveira**

**Master of Science in Software Engineering**

**2006**



**Faculty of Science and Technology  
University of Macau**

## TABLE OF CONTENTS

List of Figures .....	v
List of Tables .....	vi
List of Abbreviations .....	vii
Acknowledgments.....	viii
<b>Chapter 1: Introduction</b> .....	1
1.1 Literature review of Word Sense Disambiguation.....	2
1.1.1 Knowledge Representation For WSD.....	3
1.1.2 Knowledge based Approaches .....	5
1.1.2.1 Approaches based on Machine Readable Dictionaries.....	5
1.1.2.2 Approaches based on Thesaurus.....	8
1.1.2.3 Approaches based on Computational Lexicons.....	9
1.1.3 Corpus based Approaches.....	12
1.1.3.1 Supervised Approaches.....	12
1.1.3.2 Semi-supervised Approaches.....	17
1.1.3.3 Unsupervised Approaches .....	19
1.2 Problem Statements .....	22
1.2.1 Limited resources in Portuguese and Chinese language.....	23
1.2.2 Data Sparseness problem in the corpus .....	23
1.2.3 Word and Structural Ambiguity.....	23
1.2.4 Weaknesses in the Portuguese-Chinese Machine Translation System ...	24
1.3 Motivation.....	24
1.4 Thesis Layout and Brief Overview of Chapters .....	26
<b>Chapter 2: Sense Categorization</b> .....	29
2.1 Types of ambiguity .....	29
2.2 Sense Inventory used .....	30
2.3 Sense Types and Representation.....	30
2.4 Entries features in Sense Inventory.....	32
2.4.1 Features stored for Nouns .....	33

2.4.2 Features stored for Verbs .....	33
2.4.3 Features stored for Adjectives .....	34
2.5 Chapter Conclusion.....	34
<b>Chapter 3: Unsupervised Learning Framework.....</b>	<b>35</b>
3.1 Disambiguation Algorithm .....	36
3.1.1 Mathematical Model .....	37
3.2 Learning Framework.....	39
3.2.1 Corpus Pre-processing .....	40
3.2.2 Ambiguous and Related Words Identification.....	40
3.2.3 Sense Categorization.....	43
3.2.4 Scoring Translation Candidates .....	43
3.2.5 Rules Generation.....	44
3.2.6 Post-Editing Phase .....	46
3.3 A learning example in the proposed framework.....	46
3.4 Discussion and Comparison with the previous approaches .....	47
3.5 Chapter Conclusion.....	49
<b>Chapter 4: Extensions to the Unsupervised Learning Framework.....</b>	<b>51</b>
4.1 Data Sparseness .....	51
4.2 Use of bilingual examples in the dictionary .....	53
4.2.1 Extraction of Senses.....	54
4.2.2 Hypothesis: Consideration of Bilingual Examples as Training Corpus .....	56
4.2.3 Limitations .....	57
4.3 Application of Singular Value Decomposition techniques.....	57
4.3.1 Construction of the Multi-dimensional Vector Space .....	58
4.3.2 Singular Value Decomposition .....	59
4.3.3 Closeness Calculation .....	63
4.3.4 Advantages of dimensionality reduction in SVD .....	64
4.4 Chapter Conclusion.....	66
<b>Chapter 5: Portuguese-Chinese translation framework .....</b>	<b>68</b>
5.1 Sentence Analysis .....	69
5.1.1 Morphological Analysis.....	69

5.1.2 Part of speech tagging .....	70
5.1.2.1 Comparison between POS tagging and WSD.....	70
5.2 Word Sense Disambiguation Translation Module.....	71
5.2.1 Disambiguation based on WSD database .....	71
5.2.2 Disambiguation based on structure constituents and senses.....	72
5.2.3 Translation Disambiguation Example.....	74
5.3 Constraint based Synchronous Grammar.....	77
5.3.1 Definition of Constraint-based Synchronous Grammar .....	77
5.3.2 Parsing Constraint-based Synchronous Grammar .....	79
5.3.3 Translation Process .....	79
5.3.4 Advantages of Constraint-based Synchronous Grammar .....	81
5.4 Shallow Parsing .....	82
5.4.1 Related Chunking Approaches .....	83
5.4.2 Application of Noun Phrase Shallow Parsing in Portuguese-Chinese MT.....	85
5.4.3 Identification of Noun Phrase Chunks based on CSG grammars .....	86
5.4.3.1 Noun Phrases with Adjectives .....	86
5.4.3.2 Noun Phrases with conjunctions.....	87
5.4.3.3 Noun followed by another Noun.....	88
5.4.3.4 Determiners/Articles followed by Nouns .....	88
5.4.3.5 Noun Phrases with numberings .....	89
5.4.3.6 Pronouns followed by Nouns.....	90
5.4.3.8 Noun Phrases with Clauses.....	91
5.4.3.9 Head Nouns followed by prepositions attached with nouns or verbs.....	93
5.4.4 Advantages in the Identification of Noun Phrases.....	94
5.5 Chapter Conclusion.....	98
<b>Chapter 6: Evaluation Results</b> .....	99
6.1 Evaluation Metrics .....	99
6.1.1 Evaluation Metrics <i>in vitro</i> .....	100
6.1.1.1 Upper and Lower Bound.....	100

6.1.1.2	Applicability and Precision.....	101
6.1.1.3	SENSEVAL .....	101
6.1.1.4	Evaluation Metrics <i>in vivo</i> applied .....	102
6.2.1	Evaluation <i>in vivo</i> .....	103
6.3	Testing Environment and Resources used .....	103
6.4	Observations in the rules extracted .....	104
6.5	Evaluation Results .....	106
6.5.1	Evaluation Results <i>in vitro</i> .....	106
6.5.2	Evaluation Results <i>in vivo</i> .....	107
6.6	Chapter Conclusion.....	108
<b>Chapter 7:</b>	<b>Conclusion</b> .....	<b>109</b>
7.1	Research Contributions.....	110
7.2	Future directions .....	111
<b>References</b>	.....	<b>113</b>