



**FLINDERS
UNIVERSITY**

**ADELAIDE
AUSTRALIA**

Word Segmentation and Ambiguity in English and Chinese NLP & IR

by

黄金虎

Jin Hu Huang, *B.Eng.(Comp.Sc)Grad.Dip.(Soft.Eng.)*
School of Computer, Engineering and Mathematics,
Faculty of Science and Engineering

August 10, 2011

A thesis presented to the
Flinders University of South Australia
in total fulfillment of the requirements for the degree of
Doctor of Philosophy
in
Computer Science

CONTENTS

<i>Abstract</i>	ix
<i>Certification</i>	xii
<i>Acknowledgements</i>	xiii
<i>Preface</i>	xv
<i>1. Introduction</i>	1
1.1 Context Sensitive Spelling Correction	1
1.2 Chinese Pinyin Input	2
1.3 Chinese Segmentation	6
1.4 Chinese Information Retrieval (IR)	9
1.5 Thesis Contribution	10
1.6 Thesis Organization	13
<i>Part I Word Disambiguation for English Spelling Checking and Chinese Pinyin Input</i>	15
<i>2. Machine Learning for Context Sensitive Spelling Checking</i>	16
2.1 Introduction	16
2.2 Confused Words	17
2.3 Context-sensitive Spelling Correction	18
2.4 Experiment and Result	21
2.5 Interface	29
2.6 Conclusion and Future Work	30
2.7 Reflections	31

3. <i>Statistical N-gram Language Modeling</i>	33
3.1 Statistical Language Modeling	33
3.2 N-Gram Markov Language Models	36
3.3 Smoothing Methods	37
3.3.1 Add One Smoothing	37
3.3.2 Interpolation Smoothing	38
3.3.3 Absolute Discounting	38
3.3.4 Good-Turing Discounting	38
3.3.5 Katz Back-off Smoothing	39
3.3.6 Witten-Bell Smoothing	39
3.3.7 Kneser-Ney Smoothing	40
3.3.8 Modified Kneser-Ney Smoothing	40
3.4 Discussion	41
3.5 Conclusion	41
4. <i>Compression-based Adaptive Approach for Chinese Pinyin Input</i>	42
4.1 Introduction	42
4.2 Statistical Language Modelling	44
4.2.1 Pinyin-to-Character Conversion	45
4.2.2 SLM Evaluation	45
4.3 Compression Theory	46
4.4 Adaptive Modelling	47
4.5 Prediction by Partial Matching	48
4.6 Experiment and Result	52
4.7 Conclusion	54
5. <i>Error-driven Adaptive Language Modeling for Pinyin-to-Character Conversion</i>	55
5.1 Introduction	55
5.2 LM Adaption Methods	56
5.2.1 MAP Methods	56
5.2.2 Discriminative Training Methods	57
5.3 Error-driven Adaption	58
5.4 Experiment and Result	61
5.5 Conclusion	65

<i>Part II Chinese Word Segmentation and Classification</i>	66
6. <i>Chinese Words and Chinese Word Segmentation</i>	67
6.1 Introduction	67
6.2 The Definition of Chinese Word	67
6.3 Chinese Word Segmentation	70
6.3.1 Segmentation Ambiguity	70
6.3.2 Unknown Words	71
6.4 Segmentation Standards	72
6.5 Current Research Work	73
6.6 Conclusion	76
7. <i>Chinese Word Segmentation Based on Contextual Entropy</i>	78
7.1 Introduction	78
7.2 Contextual Entropy	79
7.3 Algorithm	80
7.3.1 Contextual Entropy	80
7.3.2 Mutual Information	82
7.4 Experiment Results	83
7.5 Conclusion	90
7.6 Reflections	90
8. <i>Unsupervised Chinese Word Segmentation and Classification</i>	92
8.1 Introduction	92
8.2 Word Classification	93
8.3 Experiments and Future Work	96
8.4 Conclusion	99
8.5 Reflections	100
<i>Part III Chinese Information Retrieval</i>	101
9. <i>Using Suffix Arrays to Compute Statistical Information</i>	102
9.1 Suffix Trees	102
9.2 Suffix Arrays	104
9.3 Computing Term Frequency and Document Frequency	109
9.4 Conclusion	112

10. <i>N</i> -gram based Approach for Chinese Information Retrieval	113
10.1 Introduction	113
10.2 Chinese Information Retrieval	114
10.2.1 Single-character-based (Uni-gram) Indexing	114
10.2.2 Multi-character-based (N-grams) Indexing	115
10.2.3 Word-based Indexing	115
10.2.4 Previous Works	116
10.3 Retrieval Models	121
10.3.1 Vector Space Model	121
10.3.2 Term Weighing	122
10.3.3 Query and Document Similarity	123
10.3.4 Evaluation	123
10.4 Experimental Setup	126
10.4.1 TREC Data	126
10.4.2 Measuring Retrieval Performance	127
10.5 Experiments and Discussion	127
10.5.1 Using Dictionary-based Approach	127
10.5.2 Statistical Segmentation Approach	128
10.5.3 Using Different N-grams	131
10.5.4 Word Extraction	136
10.5.5 Removing Stop Words	142
10.6 Discussion	145
10.7 Conclusion	147
11. <i>Conclusions</i>	149
11.1 Thesis Review	149
11.2 Future Work	151
<i>Appendix</i>	153
<i>A. The Appendix: Tables for TREC 5 & 6 Chinese Information Retrieval Results</i>	154
<i>B. The Appendix: Examples of TREC 5 & 6 Chinese Queries</i>	162
<i>Bibliography</i>	188

LIST OF FIGURES

1.1	Number of Homonyms for Each Pinyin	4
5.1	Perplexity compare between static and adaptive model on Modern Novel	62
5.2	Perplexity compare between static and adaptive model on Martial Arts Novel	63
5.3	Perplexity compare between static and adaptive model on People’s Daily 96	64
7.1	Contextual Entropy and Mutual Information for “The two world wars happened this century had brought great disasters to human being including China.”	80
10.1	Average Precision At Different Recall For Dictionary-based and Statistical Segmentation Approaches	129
10.2	Average Precision At X Documents Retrieved For Dictionary-based and Statistical Segmentation Approaches	130
10.3	Average Precision for 54 Queries Using 1-gram, 2-grams, 3-grams and 4-grams	134
10.4	The Impact of Extracted Words on 54 Queries	139
10.5	The Impact of Stop Words on 54 Queries	143

LIST OF TABLES

2.1	Diameter with occurrences, significance and probability - number of contexts (WSJ 87-89,91-92)	22
2.2	Relationship between probability and significance - number of contexts (WSJ 87-89,91-92)	23
2.3	False errors and coverage testing on test and validation corpora with no errors seeded but two real errors found	24
2.4	True errors detected (recall) and corrected when errors seeded randomly	25
2.5	Seeded errors of the confusion set of “ <i>from</i> ” and “ <i>form</i> ” ($S,P \geq 95\%$)	25
2.6	False positive rate (FPR), true positive rate (TPR) and informedness when errors seeded	27
2.7	Spelling Errors Found in WSJ0801 and WSJ1231	28
4.1	Compression models for the string “dealornodeal”	47
4.2	PPM model after processing the string <i>dealornodeal</i>	50
4.3	Compression results for different compression methods	52
4.4	Character Error Rates for Kneser-Ney, Static and Adaptive PPM	53
5.1	Witten-Bell smoothing model after processing the string <i>dealornodeal</i>	60
5.2	Comparing perplexity and CER using different smoothing methods on testing corpus	63
5.3	CER and percentage of data used for adaption	64
5.4	Testing on Xinhua 96 with different mixed models with adaption .	65
6.1	Some differences between the segmentation standards	73
7.1	Validation results based on Recall, Precision and F-Measure for Eq. 7.1 7.2 7.3 7.4	84
7.2	Validation results based on Recall, Precision and F-Measure for Eq. 7.5 7.6 7.7 7.8	84
7.3	Validation results on Recall, Precision and F-measure according to Eq. 7.9 7.10 7.11	85

7.4	Validation results based on Recall, Precision and F-Measure for Eq. 7.12 7.13 7.14	86
7.5	Results based on Recall, Precision and F-Measure on testing corpus	86
7.6	Results based on Recall, Precision and F-Measure for the Longest Forward Match Method	89
9.1	Suffixes and suffix arrays before sorting	105
9.2	Suffixes and suffix arrays after sorting	106
9.3	Nontrivial classes for string “to_be_or_not_to_be”	108
10.1	Comparison of segmentation approaches in TREC 6	119
10.2	Term Weighting in the Smart system	123
10.3	Average Precision for Dictionary-based and Statistical Segmentation Approaches	129
10.4	Average Precision for Dictionary-based and Statistical Segmentation Approaches	130
10.5	Precision Recall and Average Precision For Different Length Of N-grams	132
10.6	Precision At X Documents and R-Precision For Different Length Of N-grams	132
10.7	Title and Description of the Query 6,7,29,34,51	133
10.8	The Impact of Word Extraction on TREC	138
10.9	Paired samples <i>t</i> -test on IR results of combining extracted words (df=53)	140
10.10	Improved Query Samples	140
10.11	Term Frequency and Document Frequency of Extracted Word and its Bi-grams	142
10.12	The Impact of Removal of “UN” and “Peace-keeping troops” on Queries 15 & 16	146
A.1	Average Precision for 54 Queries Using 1,2,3,4-grams	154
A.2	The Impact of Extracted Words on 54 Queries	157
A.3	The Impact of Stop Words on 54 Queries	159

ABSTRACT

Statistical language learning is the problem of applying machine learning technique to extracting useful information from large corpus. It is important in both statistical natural language processing and information retrieval. In this thesis, we attempt to build some statistical language learning and modeling algorithms to solve some problems in both English and Chinese natural language processing. These problems include context sensitive spelling correction in English, adaptive language modeling for Chinese Pinyin input, Chinese word segmentation and classification, and Chinese information retrieval.

Context sensitive spelling correction is a word disambiguation process to identify the word-choice errors in text. It corrects real-word spelling errors made by users when another word was intended. We build large scale confused word sets based on keyboard adjacency. Then we collect the statistics based on the surrounding words using affix information and the most frequent functional words. We store the contexts significant enough to make a choice among the confused words and apply this contextual knowledge to detect and correct the real-word errors. In our experiments we explore the performance of auto-correction under conditions where significance and probability are set by the user. The technique we developed in this thesis can be used to resolve lexical ambiguity in the syntactic sense.

Chinese Pinyin-to-character conversion is another task of word disambiguation. Chinese character can't be entered by keyboard directly. Pinyin is the phonetic transcription of Chinese characters using the Roman alphabet. The process of Pinyin-to-character conversion, similar to speech recognition, is to decode the sequence of Pinyin syllables into corresponding characters based on statistical

n-gram language models. The performance of Chinese Pinyin-to-Character conversion is severely affected when the characteristics of the training and conversion data differs. As natural language is highly variable and uncertain, it is impossible to build a complete and general language model to suit all the tasks. The traditional adaptive maximum a posteriori (MAP) models mix the task independent model with task dependent model using a mixture coefficient but we never can predict what style of language users have and what new domain will appear. We present a statistical error-driven adaptive n-gram language model to Pinyin-to-character conversion. This n-gram model can be incrementally adapted during Pinyin-to-Character converting time. We use a conversion error function to select what kind of data to adapt the model. The adaptive model significantly improves Pinyin-to-Character conversion rate.

Most Asian languages such as Chinese and Japanese are written without natural delimiters, so word segmentation is an essential first step in Asian language processing. Processing at higher levels will be impossible if there is no effective word segmentation. Chinese word segmentation is a basic research issue on Chinese language processing tasks such as information extraction, information retrieval, machine translation, text classification, automatic text summarization, speech recognition, text-to-speech, natural language understanding, and so on. This thesis presents a purely statistical approach to segment Chinese sequences into words based on contextual entropy on both sides of a bi-gram. It is used to capture the dependency with the left and right contexts in which a bi-gram occurs. Our approach tries to segment text by finding the word boundaries instead of the words. Although developed for Chinese it is language independent and easy to adapt to other languages, and it is particularly robust and effective for Chinese word segmentation.

Traditionally Chinese words are not regarded being inflected with respect to tense, case, person and number, this information is captured by separate words that attach as clitics rather than affixes. Telling the part-of-speech of a word is not straightforward. In this thesis we classify Chinese words according to the substitutability of linguistic entities from the same class. We merge words/classes

together based on contextual information and class overlapping.

Traditional information retrieval systems for European languages such as English use words as indexing units and thus cannot apply directly to Asian languages such as Chinese and Japanese due to lack of word delimiters. A pre-processing stage called segmentation has to be performed to determine the boundaries of words before traditional IR approaches based on words can be adapted to Chinese language. Different segmentation approaches, N-grams based or word based, have their own advantages and disadvantages. No conclusion has been reached among different researchers as to which segmentation approach is better or more appropriate for the purpose of IR even on standard Chinese TREC corpus. In this thesis we investigate the impact of these two segmentation approaches on Chinese information retrieval using standard Chinese TREC 5 & 6 corpus. We analyze why some approaches may work effectively in some queries but work poorly in other queries. This analysis is of theoretical and practical importance to Chinese information retrieval.

CERTIFICATION

I certify that this thesis does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any university; and that to the best of my knowledge and belief it does not contain any material previously published or written by another person except where due reference is made in the text.

As requested under Clause 14 of Appendix D of the *Flinders University Research Higher Degree Student Information Manual* I hereby agree to waive the conditions referred to in Clause 13(b) and (c), and thus

- Flinders University may lend this thesis to other institutions or individuals for the purpose of scholarly research;
- Flinders University may reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Signed

Dated

Jin Hu Huang

ACKNOWLEDGEMENTS

I wish to express my sincere gratitude to my supervisor Prof. David Powers for bringing me into natural language processing research field. I thank him for his guidance, tolerance, patience, support and encouragement over the past 10 years.

I am grateful to the members of AI group and KDM group for their discussion, friendship and support. I would like to thank the school for the financial support to attend conference in Beijing and Singapore.

I thank my family for their eternal understanding, patience and most of all, support. An enormous thanks go to my wife, Li Ping, for her love and support. Going overseas, becoming a mother and being a student were tough, but she has done so well and even found a job in government. So many times I promised to her that I would take time to finish this year. Finally I will make it in 2011. I would like to thank my parents-in-law and sister-in-law, who came over from China for the tremendous help. Feel sorry to my lovely daughter and son, Lydia and Jurgen. Sometime they are my excuses not finishing on time.

Specially thanks to Mr Eric Robins who supported me to come to study at Flinders at beginning. I would also like to thank Mr Jurgen Kracht who provided accommodation for me. I really enjoyed staying with him and his two cats, Heihei and Mushi in early my study.

I would like to thank the members of Somerton Baptist Church and Flinders Overseas Christian Fellowship for their sharing, prayers and friendship. Specially Kenneth Lim, Barry Wee, Merrilyn Teague, Dahlis and Robert Willcock, John and Marrison.

Due to some personnel reasons, I studied part time since 2003 and intermitted my study several times during 2005-2010.

This work was carried out during the tenure of International Postgraduate Research Scholarship and Flinders University Research Scholarship, and also received support under an Australia Research Council Discovery grant in the early stages.

To my parents and wife!

Jin Hu Huang

July 2011

Adelaide.

PREFACE

Almost all the results presented in this thesis have already been published in national and international conferences.

Chapter 2 is based on (Huang & Powers 2001)

Huang, J. H. & Powers, D. (2001), Large scale experiments on correction of confused words, in 'Australasian Computer Science Conference', IEEE Computer Society Press, Gold Coast, Queensland.

Chapter 4 is based on (Huang & Powers 2004)

Huang, J. H. & Powers, D. (2004), Adaptive compression-based approach for Chinese Pinyin input, in 'Third ACL SIGHAN Workshop on Chinese Processing'.

Chapter 5 is based on (Huang & Powers 2011)

Huang, J. H. & Powers, D. (2011), Error-driven adaptive language modeling for Pinyin-to-character conversion, in 'International Conference on Asian Language Processing (IALP2011)', Nov 15-17, 2011, Penang, Malaysia.

Chapter 7 is based on (Huang & Powers 2003)

Huang, J. H. & Powers, D. (2003), Chinese word segmentation based on contextual entropy, in '17th Pacific Asia Conference on Language, Information and Computation', Singapore.

Chapter 8 is based on (Huang & Powers 2002)

Huang, J. H. & Powers, D. (2002), Unsupervised Chinese word segmentation and classification, in 'First Student Workshop in Computational Linguistics', Beijing, China.

Chapter 10 is based on (Huang & Powers 2008)

Huang, J. H. & Powers, D. (2008), Suffix-tree-based approach for Chinese

information retrieval, in 'International Conference on Intelligent Systems Design and Applications (ISDA)'.

Rather than including the published paper as is permitted under the PhD rules, I have sought to integrate the material into the thesis in a cohesive way whilst achieving a balance between chapter that stand alone and avoidance of redundancy in relation to literature review.