FLINDERS
UNIVERSITY

ADELAIDE
AUSTRALIA

Word Segmentation and Ambiguity in English and Chinese NLP & IR

by

黃金虎
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

Adelaide, South Australia, 2012

CONTENTS

LIST OF FIGURES

LIST OF TABLES

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 preprocessing 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 Marrilyn.

Due to some personnel reasons, I studied part time since 2003 and intermitted my study several times during 2005-1010.

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 'Australiasian 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.