摘要

本論文主要探討的問題是如何由大量文字中自動抽取新詞 (new words; unknown words) 或複合詞 (compound words)；主要重點則在於如何有系統的改善精確率 (precision) 及召回率 (recall) 的綜合績效。爲了能針對各不同語言在不同狀況下作最佳化處理，本論文分別就英文複合詞抽取及中文未知詞 (或新詞) 之抽取提出兩種不同層次的最佳化處理技術。

一般英文複合詞抽取系統，是根據候選詞的某些結合性特徵 (association features), 使用一個過濾器 (filter) 或分類器 (classifier) 判斷該候選詞爲 "詞" 或 "非詞"。對於此類系統，本論文採用一套兩階段的最佳化策略，以使精確率及召回率的綜合績效達到最大或相對最大。第一階段的主要目的在於減少分類的錯誤率，第二階段的任務則在於，由第一階段所達成的最小錯誤率狀態出發，提昇精確率及召回率的綜合績效。爲達到最小錯誤率的目標，本論文採用一系列的改善措施，有系統地改進錯誤率。同時，爲達成第二階段的任務，本論文特別發展了一套可以依據使用者所指定的精確率及召回率函數，達到最大綜合績效的非線性學習策略。在中文未知詞抽取方面，一般系統會依據前後文的資訊 (contextual information), 作中文斷詞，再將候選詞送給分類器作判斷。由於有前後文訊息可資利用，因此，除了改變分類器的參數估計值，過濾掉一些不正確的字元組，來改善精確率之外，本論文也提出一套反覆式的方法，逐步找出相對於這些錯誤候選詞的正確新詞，以同時改善召回率；而不致因提高精確率而降低召回率，或反其道而行。

在抽取英文複合詞方面，本論文所提的技術在抽取二元 (bigram) 複合詞時，精確率及召回率的加權平均值 (weighted precision-recall, WPR) 可達 88% 左右，三元 (trigram) 複合詞的 WPR 亦達88% 左右。另一綜合績效指標 F-measure，則分別爲 84% (bigram) 及 86% (trigram) 左右。(語料由 20,715 句訓練語料及 2,301句測試語料所組成)。採用本論文所提的最佳化學習技術時，亦可明顯觀察到精確率及召回率隨指定的績效指標而變動的情形。

在中文未知詞抽取方面，精確率及召回率大致呈單調遞增的趨勢。而不像一般系統，在精確率提高時召回率會降低，或發生相反情況的現象。在測試 311,591 句中文語料時，所得到的 F-measure 分別爲 76% (bigram), 54% (trigram) 及 70% (quadgram). 與採用非反覆式方法所獲的的績效 74% (bigram), 46% (trigram) 及 58% (quadgram) 相比，有相當顯著的改善。

Abstract

Automatic lexicon acquisition from large text corpus is surveyed in this dissertation, with special emphases on optimization techniques for maximizing the joint precision-recall performance. Both English compound word extraction and Chinese unknown word identification tasks are studied in order to explore precision-recall optimization techniques in different languages of different complexity using different available resources.

In the English compound word extraction task, the simplest system architecture, which assumes that the lexicon extraction task is conducted using a classifier (or a filter) based on a set of multiple association features, is studied. Under such circumstances, a two stage optimization scheme is proposed, in which the first stage aims at minimizing classification error and the second stage focuses on maximizing joint precision-recall, starting from the minimum error status. To achieve minimum error rate, various approaches are used to improve the error rate performance of the classifier. In addition, a non-linear learning algorithm is developed for achieving maximum precision-recall performance in terms of user specified objective function of precision and recall. In the Chinese unknown word extraction task, where contextual information as well as word association metrics are used, an iterative approach, which allows us to improve both precision and recall simultaneously, is proposed to iteratively improve the precision and recall performance.

For the English compound word extraction task, the weighted precision and recall (WPR) using the proposed approach can achieve as high as about 88% for bigram compounds, and 88% for trigram compounds for a training (testing) corpus of 20715 (2301) sentences sampled from technical manuals of cars. The F-measure performances are about 84% for bigrams and 86% for trigrams. By applying the proposed optimization method, the precision and recall profile is observed to follow the preferred criteria of different lexicographers.

For the Chinese unknown word identification task, experiment results show that *both* precision and recall rates are improved almost *monotonically*, in contrast to non-iterative segmentation-merging-filtering-and-disambiguation approaches, which often sacrifice precision for recall or *vice versa*. With a corpus of 311,591 sentences, the performance is 76% (bigram), 54%(trigram), and 70% (quadgram) in *F-measure*, which is significantly better than using the non-iterative approach with F-measures of 74% (bigram), 46% (trigram), and 58%(quadgram).

Acknowledgements

Firstly, I would like to express my sincere thank to my advisor, Professor Keh-Yih Su, who had initiated my interests in the field of natural language processing in general, and machine translation in particular. I had also learned a lot from him in statistical approaches for language processing. This forms the basis for me to enjoy corpus-based research for more than ten years.

I would also like to express my thanks to Ms. Emma Chang, Mr. Ming-Wen Wu, and Ms. Yu-Ling Una Hsu of the Behavior Design Corporation (BDC), who spent lots of their precious time in preparing the corpus materials, and carefully examined such materials. Their kindness and labor intensive works make it possible for me to concentrate on the modeling problems of the current work with ease. Mr. Ming-Wen Wu and Dr. Yi-Chung Lin (EE, National Tsing-Hua University) had generously shared part of their codes with me so that I don’t need to do everything from scratch. My special thank goes to the Behavior Design Corporation and the Behavior Tech Computer Corporation (BTC). They had kindly provided me with huge amount of computing resources, which make the simulation on large corpora easy to accomplish.

My greatest thanks will be given to my parents. They had spent their youth and contributed so much to our family that I cannot express my gratitude with words. I hope that they will be healthy and happy all the time. My thanks to my two elder sisters are also beyond any expression that could be speak out with words. In the old days when we are economically tight, they had voluntarily given up their opportunities for higher education; without their sacrifice, there won’t be three highly educated boys in our family. I would like to forward my best wishes to them and their families. I also want to express my deepest gratitude to my eldest brother and his gentle wife, not because of their uninterrupted financial supports throughout my undergraduate days, but because of the paradigm they have set up for me to learn. Finally, I would like to thank my second eldest brother, who initiated my consciousness to care about the weaks. I am proud of being a member of my family. This dissertation is therefore dedicated to my family...