

University of Macau

Abstract

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

by Leong Ka Seng

Thesis Supervisors: Prof. Li Yiping and Dr. Wong Fai
Master of Science in Software Engineering

Unlike English, Portuguese, etc., Chinese text is written in a continuous way without delimiters to specify word boundaries. Therefore, for any Chinese Information Processing (CIP) system like machine translation, web information retrieval, etc., the first and essential step is word segmentation which is to detect the word boundaries within sentences. Segmentation ambiguities and unknown words are the two main problems in Chinese segmentation. These make segmentation become a non-trivial task. Different segmentations for a sentence can produce different meanings. Hence, it has important influences to a CIP system as a whole.

There are different approaches in developing Chinese word segmentation algorithms. The use of corpus-based approaches for various Natural Language Processing (NLP) tasks has been the main trend in recent years. This is because of fast development of the Internet and a lot of digitized resources can be obtained from the cyber net. In this thesis, the research of word segmentation and its application to Chinese-Portuguese Machine Translation (MT) is based on this direction.

There are two subcategories of corpus-based approach: pure statistics and machine learning based methods. Based on statistics model, an integrated Chinese segmentation and tagging model, CSAT, is proposed. CSAT applies the N-Shortest-Paths (NSP) model to achieve rough segmentation purpose. The model adopts directed graphs for describing sentences where nodes in the graphs represent Chinese characters of sentences. By attaching related statistics to the edges between nodes according to the co-occurrence frequency of characters, a statistics based model

to predict the possible word segmentations is constructed. The proposed model absorbs the advantages of full segmentation and maximum probabilities, which gives a good recall rate in empirical results. In order to obtain the final results for input sentences from the candidate segmentations, CSAT assigns proper Part Of Speech (POS) information to words of the candidates. Based on this, the segmentations with maximum probabilities will be selected for the input sentences.

From the linguistic point of view, the analysis of natural language cannot escape from considering the features of language, in particular the contextual information of Chinese words. In the second part of our researches, Maximum Entropy Based Segmentation, MESEG, is proposed to integrate different linguistic features into model for the segmentation task. As an extension of CSAT model, in many cases, MESEG outperforms in recognizing the unknown words correctly. This fully illustrates the contributions of word context features. In order to maximize the performance of MESEG, the selection of features has been the vital task and is fully studied and discussed in this thesis work. From the experiment results, MESEG is shown to outperform CSAT with a significant amount of improvement in segmentation precision.

Unknown words, in particular the unknown named entities (NE), has been recognized as the main obstacle in Chinese segmentation. This problem seriously affects the accuracy of any word segmentation system in general. In response to this, Unknown Named Entity Identification (UNEI) has been one of the main research missions in this thesis, and UNEI model based on Maximum Entropy (ME) has been proposed to tackle this problem. In application to the segmentation system, the integration of UNEI model into the MESEG has been designed and proposed to recognize NEs and segment words in a whole. The empirical experiments show that the proposed model is effective to improve the accuracy of the whole word segmentation process.

Based on the research results, an applicable word segmentation module is implemented, and applied to the Chinese-Portuguese MT system. Meanwhile, a Maximum Entropy (ME) modeling module has been developed as a standalone

package, which is ready to be used for other classification problems including different NLP tasks such as string tagging and syntax recognition.