Technical Report of *Chinese Morphological Analyzer(Chen)*

Vernkin Smith

February 27, 2009

Abstract

Practical results show that performance of statistic segmentation system outperforms that of hand-crafted rule-based systems. And the evaluation shows than the accuracy drop caused by out-of-vocabulary (OOV) words is at least five times greater than that of segmentation ambiguities. The better performance of OOV recognition the higher accuracy of the segmentation system in whole, and the accuracy of statistic segmentation systems with character-based tagging approach outperforms any other word-based system. This Report is about a supervised machine-learning approach to Character-based Chinese word segmentation. A maximum entropy tagger is trained on manually annotated data to automatically assign to Chinese characters, or *hanzi*, tags that indicate the position of a *hanzi* within a word.

Contents

Chapter 1

Overview

1.1 A Decade Review of Chinese Word Segmentation

During the last decade, especially since the First International Chinese Word Segmentation Bakeoff was held in July 2003, the study is automatic Chinese word segmentation has been greatly improved. Those improvements could be summarized as following: (1) on the computation sense Chinese words in read text that have been well-defined by "segmentation guidelines + lexicon + segmented corpus"; (2) practical results show that performance of statistic segmentation system outperforms that of hand-crafted rule-based systems; (3) the evaluation shows than the accuracy drop caused by out-of-vocabulary (OOV) words is at least five times greater than that of segmentation ambiguities; (4) the better performance of OOV recognition the higher accuracy of the segmentation system in whole, and the accuracy of statistic segmentation systems with character-based tagging approach outperforms any other word-based system.

1.2 Main Difficulties in Chinese Segmentation

Notice!!! As I don't resolve the problem of inputing Chinese Characters in the latex, all the Chinese Characters are unavailable now (instead of *pinyin* and *tone*(varies from 0 to 4)).

It is generally agreed among researchers that word segmentation is a neccessary first step in Chinese language processing. However, unlike Engish text in which sentences are sequences of words delimited by white spaces, in Chinese text, sentences are represented as strings of Chinese characters or *hanzi* without similar natural delimiters. Therefor, the first step in a Chinese language processing task is to identify the sequence of words in a sentence and mark boundaries in appropriate places. This may sound simple enough but in reality identifying words in Chinese is a non-trivial problem

that has drawn a large body of research in the Chinese language processing community.

It is easy to demonstrate that the lack of natural delimiters itself is not the heart of the problem. In a hypothetical language where all words are represented with a finite set of symbols, if one subset of the symbols always start a word and another subset, mutually exclusive from the previous subset, always end a word, identifying words would be a trivial exercise. Nor can the problem be attributed to the lack of inflectional morphology. Although it is true in Indo-European languages inflectional affixes can generally be used to signal word boundaries, it is conceivable that a hypothetical language can use symbols other than inflectional morphemes to serve the same purpose. Therefore the issue is neither the lack of natural word delimiters nor the lack of inflectional morphemes in a language, rather it is whether the language has a way of unambiguously signaling the boundaries of a word.

The real difficulty in automatic Chinese word segmentation is the lack of such unambiguous word boundary indicators. In fact, most hanzi can occur in different positions within different words. The examples in Table 1 show how the Chinese character 产("produce") can occur in four different positions. This state of affairs makes it impossible to simply list mutually exclusive subsets of hanzi that have distinct distributions, even though the number of hanzi in the Chinese writing system is in fact finite. As long as a hanzi can occur in different word-internal positions, it cannot be relied upon to determine word boundaries as they could be if their positions were more or less fixed.

Table 1. A hanzi can occur in multiple word-internal positions

Position	Example
Left	产生 'to come up with'
Word by itself	产小麦 'to grow wheat'
Middle	生产线 'assemby line'
Right	生产 'to produce'

The fact that a hanzi can occur in multiple word-internal positions leads to ambiguities of various kinds.

Chapter 2

Design

2.1 Basic Architecture of the MA

2.2 Common Architecture of the CMA

2.3 Design of the Character-based Segmentation using Maximum Entropy

Chapter 3

Appendix - Project schedules and milestones

	Milestone	Start	Finish	In Charge	Description	Status
1	Survey on the project	2008-02-01	2008-02-13	Vernkin	1. Have a general overview of current Chinese Segmentation techniques and their differences. 2. Select a suitable one (Character-based) and do more research on it.	Finished
2	Design the Architecture of CMA	2008-02-16	2008-02-20	Vernkin	1. Basen on WISE KMA Orange, design the Architecture for the MA Systems (CJK). 2. Moreover, design the common Architecture for the CMA. 3. Finally focus on the Architecture for the approach I selected.	Finished
3	Implement CMA and Unit Test	2008-02-23	2008-03-13	Vernkin	1. Implement the Maximum Entropy Model. 2. Implement character-based Segmentation. 3. Integrate all the components into CMA.	30% Finished