Technical Report of *Chinese Morphological Analyzer(Chen)*

Vernkin Smith

February 6, 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

To be continued.

Chapter 2

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.	50% Finished