International Journal of Innovative Computing, Information and Control
Volume **5**, Number **12(A)**, December **2009**

 ISSN 1349-4198
pp. **4523–4530**

CHINESE LEXICAL ANALYSIS BASED ON HYBRID MMSM MODEL

Xiao Sun[1,2], Degen Huang[1] and Fuji Ren[2,3]

[1]Department of Computer Science and Engineering
Dalian University of Technology
Dalian, 116023, P. R. China
sunxiao@is.tokushima-u.ac.jp; huangdg@dlut.edu.cn

[2]Department of Information Science and Intelligent Systems
The University of Tokushima
Tokushima, 7708506, Japan
ren@is.tokushima-u.ac.jp

[3]Department of Computer Science and Engineering
Beijing University of Posts and Telecommunications
Beijing, 100876, P. R. China

Received July 2008; revised December 2008

Abstract. *In this paper, we describe a scheme for Chinese word segmentation and POS tagging which integrates the character-based and word-based information in the directed graph generated by the MMSM model. Word-level information is effective for analysis of known words, while character-level information is useful for analysis of unknown words. A Hidden semi-CRF model is proposed for the unknown words detection and POS tagging. The proposed Hidden semi-CRF has two state chains with unequal states which can perform segmentation and POS tagging of unknown words simultaneously. The hybrid model was evaluated using the test data from SIGHAN-6 and achieved higher F-score than the stage-of-the-art models.*
Keywords: Chinese morphological analysis, MMSM model, CRF, Hidden semi-CRF

1. **Introduction.** Chinese morphological analysis including Chinese word segmentation and POS (part-of-speech) tagging is a basic step in the tasks of Chinese language processing. It also presents a significant challenge since Chinese language is typically written without separations between words. Word segmentation has been the focus thus long of significant research because of its role as a necessary pre-processing phase for the many other Chinese language processing tasks. Meanwhile, the POS tagging and unknown word (also called new words) recognitions are also the basic steps in Chinese morphological analysis. For the unknown words, which are the main difficulty in Chinese morphological analysis, both the word boundaries and the POS tags of the unknown words are unknown. The Chinese morphological analysis technologies can be categorized into three types, rule-based, machine learning and hybrid model. Among them, the machine learning-based techniques showed excellent performance in many research studies [1-3]. This method treats the word segmentation problem as a sequence of word classification. The classifier online assigns either "boundary" or "non-boundary" label to each word by learning from the large annotated corpora. Machine learning based method is also adopted in other word sequence inference techniques, such as POS tagging, phrases chunking [4] and named entity recognition [5]. But there are some cost problems in such machine learning problems, and sometimes choosing between word-based and character-based is also a dilemma.