International Journal of Innovative Computing, Information and Control
Volume **6**, Number **3(B)**, March **2010**

ICIC International ©2010 ISSN 1349-4198
pp. **1471–1480**

EXPLOITING LEXICAL INFORMATION FOR FUNCTION TAG LABELING

Caixia Yuan[1,2], Xiaojie Wang[2] and Fuji Ren[1,2]

[1]Faculty of Engineering
Tokushima University
Tokushima 770-8506, Japan
{ yuancai; ren }@is.tokushima-u.ac.jp

[2]School of Computer
Beijing University of Posts and Telecommunications
Beijing 100876, P. R. China
xjwang@bupt.edu.cn

Received November 2008; revised July 2009

Abstract. *This paper proposes a novel approach to annotate function tags for unparsed text. What distinguishes our work from previous attempts is that we assign function tags directly basing on lexical information other than on parsed trees, thus our method is general and easily portable to languages in shortage of parsing resources. In order to demonstrate the effectiveness and versatility of our method, we investigate function tag assignment for unparsed Chinese text by applying two statistical models, one is log-linear maximum entropy model, another is maximum margin based support vector machine model. We show that function tag types could be determined via powerful lexical features and effective learning algorithms. Currently, our method achieves the best F-score of 86.4 when tested on the Penn Chinese Treebank data, the highest score ever reported for Chinese text.*

Keywords: Function tags, Unparsed text, Penn treebank, Chinese language processing, Machine learning

1. **Introduction.** Recent research on shallow parsing for natural language tends to concentrate on the phrase structure identification of sentences, namely chunk parsing [1], which offers a particularly promising way to get rid of complicated full syntactic analysis. However, chunk parsing deals with the recognition of partial constituent structures at the level of individual chunks, up to date, little attention has been paid to the question of how these constituents can be structured into complete utterance. In comparison, the function types such as subject, object, etc. can tell the information like "who does what" implied in sentence, which will be useful in many natural language processing applications, such as information extraction, dialog system and machine translation system.

Broad-coverage corpora annotated with functional tags, semantic roles, or argument structures, are becoming available for English Language. The Penn Treebank [2] presents sentences with phrase structures and functional tags. In addition, the Propbank project [3] and the FrameNet project [4] share the goal of documenting the syntactic realization of predicate-argument relations of general English lexicon by annotating a corpus with semantic roles. There have also been research activities of annotating function tags for other languages. Zhou et al. [5,6] have led some pioneering work in Chinese functional chunking and initiated the construction of ThCorp ChunkBank for Chinese chunk resource. Iida et al. [7] reported the work on the specifications of annotated Japanese corpus for coreference resolution and predicate-argument analysis.