UNKNOWN CHINESE COMPOSITE WORDS TAGGING USING SELECTIVE BACK-OFF SMOOTHING

Samuel W.K. Chan* **Mickey W.C. Chong**
The Chinese University of Hong Kong
Tom B.Y. Lai
City University of Hong Kong

ABSTRACT

The aim of this research is to tag unknown Chinese words with their part-of-speech (POS).[1] Even narrow coverage of unknown words produces explosive ambiguity in natural language processing. At the same time, a completely unsupervised and refined POS tagging is impossible without any help from lexicographers. In this research, we propose to implement a means of un-locking POS tags based on two important features: *word structure and word sequence in raw text*. A similarity-based technique will be employed to classify an unknown word using its orthographic form and its contextual neighbors without becoming trapped in a subjective linguistic quagmire. The technique produces a good estimate of POS tags of Chinese compound words before they are fed into a tagger. A recursive inferential mechanism is also devised to alleviate the ripple effect from changes made at its neighbors during tagging. The approach is justified with a compound words database with more than 53,500 words. Experimental results with 500,000 words show the approach outperforms its counterparts.

* Corresponding author: swkchan@cuhk.edu.hk

Linguistic Corpus and Corpus Linguistics in the Chinese Context
Journal of Chinese Linguistics Monograph Series 25 (2015): 139-159

2409-2878/2015/25-0007

KEYWORDS
Part-of-speech tagging **C**hinese word structures **M**orphemes
Machine learning

漢語未登錄複合詞的詞性標注

陳偉光 莊華祥　　黎邦洋
香港中文大學　香港城市大學

提要
本文旨在研究漢語未登錄複合詞的詞性標注。未登錄詞往往是漢語分析的難題，在計算語言學中也帶來嚴峻的挑戰。沒有詞典編纂者的協作下，要建立一個既精確及自動化的詞性標注系統差不多是一件遙不可及的事。本研究分析複合詞內部結構和詞序等信息，並透過相仿性

技術進行詞性標注。同時，本文也詳細解釋如何應用詞內部語素等特徵及詞與詞之間的上文下理關係，計算出漢語複合詞的初步詞性標注，並將這初步篩選結果輸入到詞性標注器中，以作進一步的剖析。本文也闡釋一個遞歸推理機制，以減低在剖析過程中所產生的漣漪效應。本研究建基於一個超過 53,500 個複合詞的數據庫，進行複合詞內部結構分析。同時，在一個 50 多萬詞的語料庫中進行測試。實驗結果顯示，該方法能有效地提升複合詞詞性標注的精確度。

關鍵詞
詞性標注　**漢**語複合詞內部結構　**語**素　**機**器學習