FROM CORPUS TO GRAMMAR: AUTOMATIC EXTRACTION OF GRAMMATICAL RELATIONS FROM ANNOTATED CORPUS

Chu-Ren Huang
The Hong Kong Polytechnic University

Jia-Fei Hong[*]
National Taiwan Normal University

Wei-Yun Ma
Academia Sinica, Taiwan

Petr Šimon
The Hong Kong Polytechnic University

ABSTRACT

Automatic extraction of grammatical knowledge from corpora has been one of the ultimate goals and challenges of corpus linguistics. We present in this paper [1] one of the approaches to this challenge in Chinese corpus linguistics by introducing our recent work using the Sketch Engine (SkE, also known as Word Sketch Engine)[2] platform to automatically extract grammatical relations from PoS-annotated Chinese corpora. The SkE approach requires both giga-word size corpora and comprehensive lexico-grammatical information of the language in question. On the one hand, corpus size is crucial as the automatic extraction of grammatical relations requires enough instances of the relation pairs, which in turn require an exponential jump from the million-word size corpus for observation of single lexical items. On the other hand, lexico-grammatical information is crucial to the identification of potential relational pairs based on local context. The quality of such extraction is dependent on the quality of available lexico-grammatical knowledge. We show that a comprehensive lexical grammar, based on Information-based Case Grammar (Chen & Huang 1990) and covering over 40 thousand verbs greatly help the accuracy and recall of grammatical relation detection. The paper concludes by underlining the importance of

[*] Corresponding author: jiafeihong@ntnu.edu.tw

Linguistic Corpus and Corpus Linguistics in the Chinese Context
Journal of Chinese Linguistics Monograph Series 25 (2015): 192-221

2409-2878/2015/25-0010

integrating existing grammatical information to meet the challenge of automatic extraction of grammatical knowledge from large corpora.

KEYWORDS

Mandarin Chinese **G**rammatical knowledge **A**utomatic extraction
Lexical grammar **S**ketch engine

語料與語法：標記語料中語法關係的自動抽取

黃居仁 香港理工大學
洪嘉馡 臺灣師範大學
馬偉雲 中央研究院,臺灣
石穆 香港理工大學

提要

從標記語料庫中自動抽取語法知識，一直是語料庫語言學的終極目標挑戰。本研究的研究方法，是透過已經標示詞性的中文語料庫，使用速描引擎 (Sketch Engine, SkE)平台進行自動抽取中文詞彙，以及語言的綜合詞彙語法的訊息。一方面，語料庫的大小攸關著語法關係自動抽取時，所需要的各種關係的足夠實例，這是需要從千萬字語料庫規模才能觀察得到。另一方面，詞語語法訊息是極為重要的，這是基於所屬語境的潛在關係組的辨識。自動抽取的技術品質是依靠可用詞語語法訊息的品質。我們呈現廣泛詞語語法，基於信息語法(Chen and Huang 1990)和覆蓋率超過40000個動詞，才能有效幫助句法關係偵測，進行檢測的準確度和召回率。最後，本研究強調整合現有的合理語法信息，以滿足從大型語料庫自動抽取語法知識的挑戰的重要性。

關鍵詞

漢語　語法知識　自動抽取　詞彙語法　速描引擎