CHINESE CCGBANK CONSTRUCTION FROM TSINGHUA CHINESE TREEBANK

Chang-ning Huang
Tsinghua University, Beijing
Yan Song*
Microsoft Corporation, Redmond

ABSTRACT

For the purpose of in-depth text processing in the application of natural language processing, deep grammars require to be introduced into syntactic annotation in treebank construction. Among all of the deep grammars that can provide us deep analysis of texts, Combinatory Categorial Grammar (CCG) is an effective one with type-driven lexicalized formalism and transparent interface between syntax and semantics. In this paper, we proposed an approach of CCGbank construction based on a translation from Tsinghua Chinese Treebank (TCT). [1] In the approach, we designed a verb sub-categorization algorithm and pre-defined several Chinese sentence patterns incorporated with the standard translation procedure. Finally, the resulted CCGbank includes 32,737 sentences with more than 350,000 word tokens.[2] Evaluating experiments on both macro statistics and manually annotated references have proved the robustness of our CCGbank and the efficiency of the proposed translation process.

KEYWORDS

Combinatory categorial grammar **CCG**bank **TCT**bank **C**ategory **C**ombinatory rules

* Corresponding author: yansong@microsoft.com

Linguistic Corpus and Corpus Linguistics in the Chinese Context
Journal of Chinese Linguistics Monograph Series 25 (2015): 274-311

2409-2878/2015/25-0014

从清华中文短语结构树库到组合范畴语法树库

黄昌宁　　**宋彦**
清华大学，北京　微软，雷德蒙

提要
为了适应自然语言处理任务中的深层次文本分析，构建各类树库资源过程中需要引入深层语法以丰富其句法标注信息。在各类深层语法中，组合范畴语法（Combinatory Categorial Grammar, CCG）是一种类型驱动并高度词例化的语法，同时兼顾句法和一定程度语义信息的表达，可有效支持深层次文本分析任务。为构建具有一定规模的 CCG 资源，本文提出了从清华短语结构树库（TCTbank）自动转换得到 CCG 树库的方案，并在转换过程中使用了我们提出的一套动词次范畴化（Verb sub-categorization）以及预定义的各类中文句型转换算法，得到一个包含 32737 句，超过 35 万词次的中文 CCG 树库。该树库的可靠性以及我们采用的转换方法的有效性均通过手工和自动评价得到了验证。

关键词
组合范畴语法　**CCG** 树库　**TCT** 树库　**范**畴　**组**合规则