Mo Shen（河原大輔准教授）

「Exploiting Vocabulary, Morphological, and Subtree Knowledge to Improve Chinese Syntactic Analysis」

（語彙的、形態的、および部分木知識を用いた中国語構文解析の精度向上）

平成 28 年 3 月 23 日授与

Syntactic analysis in Chinese, including word segmentation, part-of-speech (POS) tagging, and constituency and dependency parsing, has been actively studied since the publish of the first version of Penn Chinese Treebank. However, evaluations in major natural language processing (NLP) tasks suggest that the performance of state-of-the-art systems for Chinese is constantly lower than that for European languages. This is largely due to the ambiguous nature of the definition of "word", as well as the lack of morphological inflection in Chinese, which have caused inconsistency and data sparseness problems in existing Chinese treebank annotation.

To address these difficulties, this thesis investigates the strength and weakness of previous studies and proposes new methods that improve the state-of-the-art systems. A consistent set of annotation guidelines for Chinese word segmentation, POS tagging, and dependency labelling is proposed (Figure 1.a); an algorithm that extracts substrings as reliable word boundary indicators which significantly enhance the accuracy of word segmenters is designed and implemented (Figure 1.b); a tagset for character-level POS tagging is proposed, based on which the entire Penn Chinese Treebank 5.0 is annotated (Figure 1.c); a model that performs character-level POS tagging jointly with word segmentation and word-level POS tagging is proposed (Figure 1.d); and a parse reranking model which takes advantage of global subtree features with less restriction in the structure and the size of the subtree context is proposed (Figure 1.e). All these components are integrated in a single Chinese syntactic analysis system, which is demonstrated to be effective through comprehensive parsing and machine translation experiments.

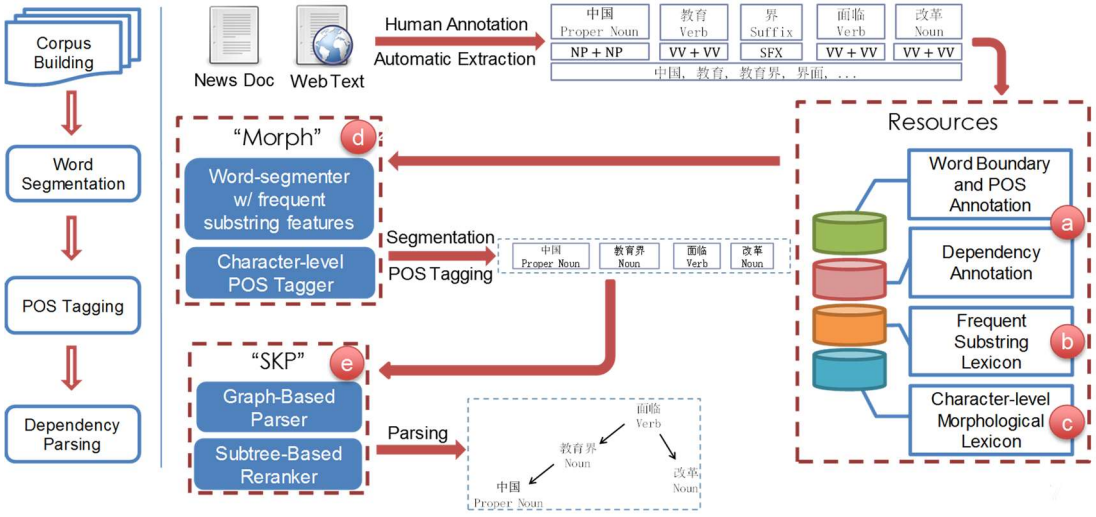

Figure 1. Illustration of the proposed system.