論文の内容の要旨

A NEW JOINT APPROACH TO WORD SEGMENTATION — INTEGRATING TASK-BASED OPTIMIZATION AND GLOBAL MORPHOLOGICAL/SYNTACTIC INFORMATION

（結合モデルによる単語分割 — タスクに基づいた最適化と大域的な形態論・統語論的情報の統合）

氏名　　羽鳥　潤

In processing natural languages that do not include delimiters (e.g. spaces) between words, word segmentation is the crucial first step that is inevitable and required to perform virtually all NLP (natural language processing) tasks, including syntactic parsing, machine translation, and information retrieval. By exploiting large corpora and machine learning methods, modern word segmentation models have achieved more than 98% accuracy in most major languages. However, there are two aspects that have not received much attention and but should not be neglected. The first is regarding the task setting itself. In a common approach to word segmentation, we fix the segmentation criterion in advance, by using a given set of rules or a dictionary, and then train a segmentation model in accordance with the given criterion. However, in non-segmenting languages, such as Japanese, Chinese, and Thai, there exists no agreement on the question: “What is a word?”, for the segmentation criterion is grammar/dictionary-dependent and essentially subjective. The second is regarding the segmentation model used. Although most of the state-of-the-art systems rely only on local context to resolve segmentation ambiguity,

there is a significant amount of ambiguity that cannot be correctly processed without considering global morphological/syntactic information.

In this thesis, by focusing on the above-mentioned two problems, we aim to reconsider the paradigm of the traditional word segmentation framework. Specifically, we propose to use joint approaches in two different manners. First, instead of using a given segmentation criterion, we propose to use the task-based optimization of segmentation units. If the segmentation is merely an intermediate representation to produce a task-specific output, you do not necessarily need to follow a given (e.g. dictionary-defined) segmentation criterion, but instead can optimize the segmentation itself so that it optimizes the quality of the final output. In Japanese pronunciation prediction task, considering a larger unit of words (e.g. compound nouns) is useful to capture broader context, while character/morpheme-level information within a word is also necessary to predict the pronunciation of out-of-vocabulary (OOV) words. By considering various word units simultaneously and allowing the model to choose the best segmentation unit among them, we show that our joint model has succeeded in predicting pronunciations of both dictionary words and OOV words within a single framework, also improving in accuracy. Second, instead of solving the task of segmentation in isolation, we argue that the word segmentation be solved along with morphological and syntactic analysis. We proposed a joint model that simultaneously processes word segmentation, morphological analysis, and syntactic parsing, and trying to capture global interaction among these three tasks in an effective way. The architecture of our model is based on an incremental parsing framework, which has an advantage in computational efficiency compared to previous works. Empirical results on Chinese treebanks show that the use of the syntactic dependency significantly improves the segmentation and POS tagging accuracy, particularly that for out-of-vocabulary (OOV) words. Also, the task of dependency parsing is shown to be significantly improved because of the relieved error propagation problem.