論文の内容の要旨

STUDY ON FEATURE GENERALIZATION FOR NAMED ENTITY RECOGNITION
（固有表現抽出のための素性の一般化の研究）

氏名　　趙　漢哲

Information extraction (IE) has been regarded as an important research issue in in the field of natural language processing (NLP). It aims to extract useful information from various types of text such as web blogs, e-mails, newswire articles, and research papers. Extracted information can be directly consumed by human users or integrated into knowledge bases that are easily accessible by other applications. For extracting information from text, it is essential to recognize smallest information units that participate in the construction of more complex knowledge. These information fragments are called named entities and the task of extracting named entities is named entity recognition (NER).

While many solutions have been proposed from rule-based to statistical approaches, most recent state-of-the-art NER systems are based on supervised learning techniques that use manually annotated data for training. However, preparing annotated data for a target domain is time-consuming and costly work; as a result, the amount of training data is often very limited. In previous studies, the data sparseness problem, which mainly results from the small size of training data, has been considered as a major obstacle in supervised learning approaches.

In this thesis, we address this problem in two perspectives. First, we propose a feature generation method that incorporates multiple segment representations (SRs) such as IOB2 and IOBES into a single model. This method enables a model to exploit not only sophisticated features that can capture the characteristics of words that often appear at specific positions but also robust features that are not much sensitive to specific positions of entity words. Second, we propose to use a new type of resource, a context gazetteer, which consists of syntactic patterns with which entity mentions can co-occur. Unlike previous studies, we build a context gazetteer from a large encyclopedic database to gather rich and sophisticated context patterns.

To investigate the effect of the proposed feature generation method, we applied it to the BioCreative 2 gene mention recognition (BC2GMR) task and the CoNLL 2003 NER (CoNLL2003) shared task. In case of traditional NER models using only one SR, a model using a more complex SR achieves higher precision than those using less complex SRs, whereas its recall starts to drop when the SR becomes too complex. On the other hand, the models using multiple SRs improve both precision and recall as more and more SRs are incorporated. To evaluate the effectiveness of a context gazetteer, we applied the context gazetteer built from the EntrezGene database to the BC2GMR task. The results improve both precision and recall; and, the major improvement comes from recall. We analyze the results to show that the context gazetteer built from a large amount of unlabeled data can provide useful context features that are not easily obtainable from a small amount of manually annotated data or human curated resources.