論文の内容の要旨

Protein Coreference Resolution for Biomedical Literature

（タンパク質名に対する照応解析）

氏名　グエン ルー トゥイ ガン

（本文）Coreference resolution has long been recognized as an important component of information extraction from literature. For biomedical domain, it is also one of the lessons from BioNLP Shared Task 2009 (BioNLP-ST 2009), which was one of the biggest community-wide efforts for textmining, that coreference structures substantially hinder the progress of fine-grained information extraction. While most of the previous works on coreference resolution were concentrated on the news domains, only a few works were carried for biomedical domain, only in a small scale.

To address the problem systematically, first we studied the domain difference of coreference phenomena in newswire domain and biomedical domain through a series of corpus analyses and experiments of coreference resolution for pronouns. Our study revealed several significant differences between the two domains. For example, while gender and person features are quite useful for coreference resolution for news texts, they have no role in the bio domain where the majority of pronouns are third person and neutral gender pronouns. The differences are mostly affected by the type of entities of interest in the two domains; while the entities of interest in newswire domain are mostly persons, companies, and so on, it is biomedical entities, e.g. proteins and cells, in biomedical domain. Considering the significant difference of the two domains, it is necessary to have a task definition

that is designed for the biomedical domain rather than replicating the same task definition defined for newswire domain. With this motivation, we defined the protein coreference task and developed necessary resources, e.g. a corpus with coreference annotation, performance evaluation metrics, and an automatic evaluation system.

The protein coreference resolution task was arranged in BioNLP-ST 2011 as a supporting task. As the final results, it received participation from six groups, among which the winning system showed the performance of finding the antecedents of anaphoric protein references at the precision of 73 percent but at the recall of 22 percent.

The analysis on the results of the shared task showed many remaining problems for improvements, among which it is recognized that semantic information is one of the key factors to improve the performance. We implemeneted a coreference resolution system incorporating semantic information specific to the bio domain. Experimental results show the use of semantic information improves the performance significantly, showing 51.3 percent of f-score, which outperforms the winning system of the shared task by 17.2 percent.