論文の内容の要旨

論文題目　Relation Information Extraction Using Deep Syntactic Analysis
（深い統語解析を用いた関係情報抽出）

氏名　薬師寺 あかね

There has been an increasing need for natural language processing technology to Information Extraction (IE), such as relations between entities, which are more informative than mere documents searched by key words.
This dissertation proposes a novel method to construct and utilize extraction patterns for relation extraction based on deep syntactic relations obtained by full parsing.
The process which requires the most amount of manual work in construction of IE systems is construction of ¥textsl{extraction patterns} which extract target information from source texts, because the same information can be represented through many kinds of syntactic variations.
To reduce this amount of manual work, our approach has two phases:
First, we raise representation ability of extraction patterns and reduce number of necessary patterns by normalizing syntactic variations into predicate-argument structures (PASs) using a full parser based on Head-driven Phrase Structure Grammar (HPSG).
Then, PASs which connect entities to extract in a small training corpus are considered as extraction patterns, and we divide them into components and utilize combinations of the components for generalization.
As a real world application, we have constructed an IE system for protein-protein interactions, which are important knowledge in biomedical research.
We evaluated the IE system on a small test-case corpus and a large real-world corpus, and show its effectiveness.

This dissertation also describes aspects that should be considered to ensure effectiveness of full parsers on domain-specific IE.
The first aspect is the ability of deep syntactic relations obtained by parsing to capture syntactic information, which is necessary for constructing extraction patterns.
To show enough accuracy of full parsing on a biomedical text, we evaluated precision of primitive PASs obtained from a biomedical text by an HPSG parser. And to compare performance of PAS patterns to patterns of part-of-speeches, we also evaluated performance of verb-argument relations obtained from a biomedical text by an HPSG parser and by patterns of part-of-speeches.
The second aspect is difficulties to apply general parsers to domain-specific domains.
To measure domain-specific coverage of a general-purpose HPSG, we investigated deficiencies of the grammar on parsing a biomedical text.
We also show preliminary investigation on performance of general-purpose parsers that suggested parsing accuracy on general corpus does not ensure parsing accuracy or IE accuracy on a domain-specific text.

Through all results on this dissertation, we show that full parsing is effective for IE.
To obtain more performance of an domain-specific IE with full parsing, we should use shallow information in sentences, such as surface words, in combination of full parsing results.
And it is also necessary to develop a full parser not only with consideration to general-purpose corpora but also with consideration to domain-specific text.