# 論文の内容の要旨

論文題目

Relation Extraction from Web Contents with Linguistic and Web Features
（言語分析及び Web 上の情報を用いたコンテンツからの関係の抽出）

顔　　玉蘭

（本文）

With the advent of the Web and the explosion of available textual data, interest in techniques for machines to understand unstructured text has been growing. Recent attention to map textual content into a structured knowledge base through automatically harvesting semantic relations from unstructured text has encouraged Data Mining and Natural Language Processing researchers to develop algorithms for it. The relations can be defined in various levels regarding to their closeness to human understanding. One kind of relations is defined from the view of natural language understanding which is going through syntactic parsing towards semantic parsing. Many efforts have been focusing on how to represent sentence in structured representation. Identification of information from sentences and their arrangement in a structured format to be used in NLP and Web mining applications such as web searching and information extraction are expected. Another kind of relations is defined as binary relationships between named entities such as *birth date*, *CEO* relations. Many recent efforts in this view have been focused on harvesting large scale of relational information from a local corpus or use the Web as corpus to build semantic repositories or ontologies for different applications such as question answering, semantic search.

In the first part of this thesis, we present a shallow semantic parser to add a new layer of semantic annotation of natural language sentences, facing the challenge of extracting a universal set of semantic or thematic relations covering various types of relations to represent sentence in a uniform structured representation. Our parser is based on the Concept Description Language for Natural Language (CDL.nl) which defines a set of semantic relations to describe the concept structure of text. In the second part, we propose several relation extraction methods to extract semantic relations from Wikipedia. Currently frequent pattern mining-based methods and syntactic analysis-based methods are two types of leading methods for semantic relation extraction task. Using respective characteristics of Wikipedia articles and Web corpus, with a novel view on

integrating syntactic analysis on Wikipedia text with redundancy information from the Web, we learn to discover and enhance relations in which a specified concept in Wikipedia participates with the complementary between the Web view and linguistic view. On the one hand, from the linguistic view, linguistic features (syntactic/dependency features) are generated from linguistic parsing on Wikipedia texts by abstracting away from different surface realizations of semantic relations. On the other hand, Web features (co-occurrence relational terms/textual patterns) are extracted from the Web corpus to provide frequency information by using a search engine.

In this thesis, we report evaluation results to illustrate the effectiveness and efficiency of our methods. For our shallow semantic parser, experiments on a manual dataset show that CDL.nl relations can be extracted with good performance. For our relation extraction systems from Wikipedia, evaluations demonstrate the superiority of the view combination over existing approaches. Fundamentally, we study the interrelated connection between linguistic and web views for semantic relation extraction. Our methods demonstrate how deep linguistic features contribute complementarily with Web features to the generation of various relations. Our study suggests an example to bridge the gap between Web mining technology and “deep” linguistic technology for information extraction tasks. It shows how “deep” linguistic features can be combined with features from the whole Web corpus to improve the performance of information extraction tasks. And we conclude that learning with linguistic features and Web features is advantageous comparing to only one view of features.