# Text Structure Analysis Methods and Application to Paraphrase Identification

by

Ngo Xuan Bach (s1120004)

School of Information Science

Japan Advanced Institute of Science and Technology

March, 2014

## Abstract

Analyzing structures of texts is important to understand natural language, both general texts and texts in some specific domains such as the legal domain. For general texts, discourse structures have been shown to have an important role in many natural language processing applications, including text summarization, question answering, information presentation, dialogue generation, and paraphrase extraction. In the legal domain, where legal texts have their own specific characteristics, recognizing logical structures in legal texts does not only help people in understanding legal documents, but also to support other tasks in legal text processing.

In this thesis, we study the structures of texts based on relations between discourse units. Regarding relations between discourse units, we focus on general semantic relations and on logical relations, which are appropriate in some cases such as laws. For general semantic relations, we study a model based on Rhetorical Structure Theory (RST). For logical relations, we study a model for legal paragraphs. Both models are based on the same framework, which consists of two steps, *Recognizing discourse units of texts* and *Building structures of texts from the discourse units*.

In our work on learning discourse structures, we propose an Unlabeled Discourse parsing system in the RST framework (UDRST). UDRST consists of a segmentation model and a parsing model. Our segmentation model exploits subtree features to rerank the N-best outputs of a base segmenter, which uses syntactic and lexical features in a Conditional Random Field (CRF) framework. The advantage of our model is that subtree features are long distance non-local features which can capture whole discourse units. In the parsing model, we introduce an incremental algorithm for building discourse trees. The algorithm builds a discourse tree for each sentence, then for each paragraph, and finally for the whole text. We also propose a new algorithm that exploits the dual decomposition method to combine a greedy model and the incremental model. Our system achieves state-of-the-art results on both the discourse segmentation task and the unlabeled discourse parsing task

on the RST Discourse Treebank corpus.

Concerning our study on analyzing logical structures of legal texts, we propose a two-phase framework for analyzing logical structures of legal paragraphs. In the first phase, we model the problem of recognizing logical parts in law sentences as a multi-layer sequence learning problem, and present a CRF-based model to recognize them. In the second phase, we propose a graph-based method to group logical parts into logical structures. We consider the problem of finding a subset of complete subgraphs in a weighted-edge complete graph, where each node corresponds to a logical part, and a complete subgraph corresponds to a logical structure. We propose an integer linear programming formulation for this optimization problem. We also introduce an annotated corpus for the task, the Japanese National Pension Law corpus, and describe our experiments on that corpus.

We then study how to exploit discourse structures for identifying paraphrases. By analyzing paraphrase sentences, we found that discourse units are very important for paraphrasing. In many cases, a paraphrase sentence can be created by applying several operations to the original sentence. Motivated by the analysis of the relation between paraphrases and discourse units, we propose a new method to compute the similarity between two sentences. Unlike conventional methods, which directly compute similarities based on sentences, our method divides sentences into discourse units and employs them to compute similarities. We apply our method to the paraphrase identification task. Experimental results on the PAN corpus, a large corpus for detecting paraphrases, show the effectiveness of using discourse information for identifying paraphrases.

**Keywords**: Text Structure Analysis, Legal Text Processing, Discourse Structure, Rhetorical Structure Theory, Logical Structure, Paraphrase Identification, Discourse Unit, Text Similarity, Conditional Random Fields, Support Vector Machines.