氏　　名	NGO XUAN BACH			
学位の種類	博士(情報科学)			
学位記番号	博情第 292 号			
学位授与年月日	平成 26 年 3 月 24 日			
論文題目	Text Structure Analysis Methods and Application to Paraphrase Identification (テキスト構造解析法とパラフレーズ識別への適用)			
論文審査委員	主査	島津　明	北陸先端科学技術大学院大学	教授
		東条　敏	同	教授
		Ho Tu Bao	同	教授
		白井　清昭	同	准教授
		徳永　健伸	東京工業大学	教授

論文の内容の要旨

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.

論文審査の結果の要旨

本論文は，テキスト構造の解析法及びパラフレーズ同定の応用を述べている．自然言語処理システムは，テキスト構造を正しく捉えることにより，テキストが表す情報を正しく捉え，首尾一貫したテキストや可読性の高いテキストを生成することができる．従来の多くの研究が 1 文単位の構文主体の処理であるのに対し，本論文は，テキスト構造に着目し，一般テキスト及び法令テキストを対象に，機械学習によりテキスト構造を解析する方法を提案し，実験により有効性を示している．応用として，パラフレーズの同定問題に取り組み，新手法を提案し有効性を示している．

一般テキストの構造解析については，RST という談話構造理論に基づく新方法を提案している．RST は，談話単位と呼ばれる単位によりテキスト構造（談話構造）を捉える枠組みとして広く知られる理論であるが，その談話構造を精度よく解析する方法はこれまで知られていない．本論文は，どの談話単位がどの談話単位と関係するかという点に焦点を当て，新方法を提案している．先ず，CRFs により談話単位に分割し，構文木を素性とする reranking 法により分割精度を上げている．次に，二つの構文解析アルゴリズムを組み合わせて談話構造を解析する新方法を提案し，新方法の精度がより高いことを実験により示している．

法令テキストの構造解析については，条項のテキスト構造（要件効果構造）を解析するユニークな解析法を提案し，国民年金法を対象にした実験により有効性を示している．条項は一般に複数の文からなり，複数の要件効果構造を表している．複数の要件効果構造は各文と複雑な対応関係にある．本論文は，このような要件効果構造を 2 段階で解析する方法を提案している．先ず，要件や効果となる要素を CRFs により認識し，次に，各要素を頂点，関係する要素間を重み付き辺としてグラフ表現し，機械学習と整数計画法により適切な複数の要件効果構造を求めるというものである．

テキスト構造解析の応用として，二つの文の談話単位を比較してパラフレーズ関係を識別する新方法を示している．この方法は，要約，機械翻訳，質問応答，剽窃検知など様々な応用に用いられる技術となるもので，最先端の性能であることを実験により確認している．この論文は，国際会議発表の支援を受けた NEC C&C 財団から最優秀論文賞を受賞している．

以上，本論文は，テキスト構造の解析法について新手法と有効性を示し，学術的に貢献するところが大きい．よって博士（情報科学）の学位論文として十分価値あるものと認めた．