氏　　名	NGUYEN, Minh Hai			
学位の種類	博士(情報科学)			
学位記番号	博情第 289 号			
学位授与年月日	平成 25 年 12 月 20 日			
論文題目	A Study on Support-Sentence Retrieval (与えられたトピックを支持する文の検索に関する研究)			
論文審査委員	主査	白井　清昭	北陸先端科学技術大学院大学	准教授
		島津　明	同	教授
		東条　敏	同	教授
		Nguyen Minh Le	同	准教授
		新納　浩幸	茨城大学	准教授

論文の内容の要旨

In this thesis, we introduced a new task, namely Support-Sentence Retrieval (SSR hereafter), that is centered on sentence retrieval. The goal of a SSR system is to retrieve sentences relevant to a given theme (called support-sentences), then classifies them into relevant types such as agreement and contradiction. It would help users to write an article about the theme by giving a comprehensive view of those support-sentences. Our study is the first attempt to develop such a kind of system. The system is divided into two main modules: sentence retrieval and sentence classification.

Sentence retrieval is the task of retrieving relevant sentences against a query. It has been found useful in many tasks such as question answering, summarization, information extraction, etc. Sentence retrieval is usually considered a special case of document retrieval. In fact, the state of the art sentence retrieval method, TF-ISF, is an adaptation of document retrieval method, TF-IDF, at sentence level. However, TF-ISF relies totally on the lexical statistics (term frequency) in the collection. In our system, a full sentence is used as query. Thus, we can utilize the syntactic structure of a sentence in the retrieval. In this task, we propose a method that can utilize both the lexical and grammatical information of a query sentence. In addition, a new query term weighting scheme based on the specificity of the terms is proposed and combined with ordinary IDF weighting for a better performance. Experimental results indicate that our best configuration of sentence retrieval system achieves 32.73% higher precision than the traditional TF-ISF method.

The key idea of sentence retrieval is finding the matching clues between user's query and candidate sentences. However, extracting lexical information (e.g. query terms, n-gram) for the matching faces the problem of vocabulary mismatch due to the fact that a sentence is

very short in comparison with a document. There is very few terms that can be matched. An approach to address this problem is query expansion. Our research resort to lexical expansion (expanded terms are query-related, e.g. synonym, hypernym). However, this can introduces noise to the system that leads to error in the later processes. Therefore, in this thesis, we would also investigate on how to integrate a word sense disambiguation (WSD hereafter) classifier into sentence retrieval system. This is the first attempt to study the impact of WSD in retrieving relevant sentences. We showed that at the moment, due to the limitation of the context information that we can extract from a sentence, a supervised WSD classifier could not predict word sense effectively. In the cases that it is able to identify the correct senses, it is still difficult for the SR system to find the matched terms between query and candidate sentence although we have already removed the noise added by incorrect expanded terms.

A SSR system extracts sentences relevant to a given topic and put them into meaningful categories, such as agreement and contradiction. Previous researches have already considered the semantic relations between sentences. For example, Recognize Textual Entailments (RTE) task identifies whether the meaning of a text can be inferred from another text; Cross-document Structure Theory (CST) recognizes 18 semantic relations between sentences across topically-related documents. However, most previous approaches applied supervised learning methods which require hand-tagged data. In the sentence classification task of SSR system, we present new sentence classification algorithms based on rules and bootstrapping method to recognize two semantic categories: agreement and contradiction. The initial seed data for training bootstrapping-based classifiers are automatically built. Our best configuration of bootstrapping-based classifiers yields 2.9% higher result than the word overlap baseline in the agreement category. Applying bootstrapping learning even increases the precision at 10 (P@10) of the contradiction class by 12.1% comparing to the rule-based approach. These results are promising due to the fact that the whole process requires no human interaction.

論文審査の結果の要旨

本論文は、あるトピックに関連する文をクエリとして入力とし、そのトピックの内容を支持する文もしくは支持しない文(以下、まとめてサポート文と呼ぶ)を文書集合から検索する新しい課題を提唱し、それをシステムとして実現する方法について論じている。サポート文検索システムのうち、同論文では特にサポート文検索タスクとサポート文分類タスクに焦点を当てている。

サポート文検索タスクでは、文書集合内のサポート文の候補となる文について、クエリ文との類似度を計算し、類似度の高い文を検索結果として出力する。論文で提案されている文間の類似度計算方法は、語彙に基づく類似度と構文的関係の類似度の両方が考慮されているという特徴がある。また、両者の最適な比重を実験的に検証している。さらに、文間類似度の計算の際には特に重要な語に大きな重みを与えている。語の重みを計算する際に、単語の特殊性、すなわち単語がどれだけ(抽象的ではなく)具体的な意味を持っているかを考慮した手法を提案している。評価実験の結果、語彙的類似度と構文的関係の類似度の両方が文検索の精度向上に貢献すること、単語の特殊性を考慮した語の重み付け手法が有効であることを確認した。

上記に加えて、語義曖昧性解消のための分類器を機械学習し、それを文検索モジュールに組み込むことで、サポート文検索タスクにおける語義曖昧性解消の有効性について調査した。クエリ文は文長が短いために語義曖昧性解消の精度が低くなること、提案システムでは文間類似度の計算の際に構文的関係を考慮することで語義曖昧性解消と同様の効果が得られることなどを明らかにした。

サポート文分類タスクについては 2 つの新しい手法を提案している。ルールベースの手法では、2 つの文が同じ意味もしくは反対の意味を持つかの極性、ならびに 2 つの文の関連性により、サポート文の候補がクエリ文に対して同義もしくは反義を持つかを分類する。ブートストラップに基づく手法では、まずヒューリスティクスによってシードと呼ばれる同義または反義を持つ文を少量生成する。次に、シード文と近い文を新たに分類し、それらを同義文もしくは反義文集合に加える。上記の操作を繰り返すことで、多くのサポート文が得られる。従来の文分類の関連研究の多くが教師あり機械学習に基づく手法であるのに対し、提案手法では訓練データとしてあらかじめ人手で大量の正解データを用意しなくてもよいという大きな利点がある。

以上、本論文はサポート文検索システムという新しい研究課題に取り組み、サポート文検索および分類タスクについて優れた手法を提案したものであり、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。