Title	与えられたトピックを支持する文の検索に関する研究
Author(s)	NGUYEN, Hai Minh
Citation	
Issue Date	2013-12
Type	Thesis or Dissertation
Text version	ETD
URL	http://hdl.handle.net/10119/11932
Rights	
Description	Supervisor:白井　清昭, 情報科学研究科, 博士

Abstract

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.

Keywords: sentence retrieval, sentence classification, information retrieval, unsupervised classification, bootstrapping learning