Abstract

Semantic search (a content-based search method) is an expected direction of information retrieval in the future. Semantic search aims at finding information that is more relevant to the given questions than that of keyword-based search (which is the method of current search engines). Despite the improvements of the current semantic search systems, they still return results in form of pages. Against such page-based approaches, there are researches for directly answering information or sentences to questions instead of returning pages. For this approach, in this research, we put our attention on using semantic relations between terms or between text spans in order to appropriately answer some types of questions. Especially, in this study, we take up named-entity-related relations, and Rhetorical Structure Theory (RST) relations as semantic relations, since these relations can be clues for answering some typical question types such as *who*, *list*, *why*, *suggestion* and *how to*, because we can see abundant such relations in sentences and relate such relations to the question types.

Questions can be roughly classified into named-entity-related (NE-related) and non-named-entity-related (non-NE-related) types. NE-related relations can be clue for answering NE-related questions, and RST relations can be clue for answering non-NE-related questions. The goal of this dissertation is to exploit such relations in texts for extracting answers to some types of questions. Concretely, the following problems are targeted in this research:

- Since Named Entities (NEs) are important in many Natural Language Processing (NLP) applications including semantic search, where the queries related to named entities take a relatively large portion. Our first issue is to extract some of named-entity-related relations, then to utilize these relations for answering some named-entity-related question types.
- There are a lot of question types besides named-entity-related questions. The second issue is to exploit RST relations among text spans for extracting answers to some other question types. We exploit the assumption that, in a relation between two text spans, one text span can be the answer to a question related to the other.
- In order to make a semantic search system understand questions, given a question, it needs to identify question type to select appropriate relations as clue for extracting answers. The last issue is question classification (QC). For the issue, increasing the classification performance based on machine learning is a promising approach. Since labeled questions are expensive, while unlabeled questions are abundant and cheap to collect, we originally propose to use labeled and unlabeled questions in semi-supervised approach to improve the performance of question classification. We also propose to use hierarchy of classifiers in order to reduce the number of question classes per classifier, since when the number of question classes of a classifier is big, its performance is affected. Different learning algorithms for classifiers in the classifier hierarchy are also investigated.
- Finally, based on the above issues, we build a prototype of a semantic search system. Our system can get a question in form of an English sentence from a user, and classify the given question to carry out the corresponding answer extraction method. If found, the answer in form of a sentence or paragraph is returned to the user.

In summary, the thesis focuses on exploiting the semantic relations in documents for extracting answers to questions. The solutions for the analyzed, investigated problems have been provided. The contributions of this thesis include two aspects: the theoretical study of developing algorithms for extracting named-entity-related relations, question classification, and empirical study in extracting answers to questions. Our experiments give promising results.