Survey on Studies of Inference in Question Answering Systems Handling Reason and Implicature

Yasutoshi Yagiu (1210755)
School of Information Science,
Japan Advanced Institute of Science and Technology

August, 2014

Keywords: Why-question answering system, Recognizing textual entailment, Inference, Natural language processing

With progress of information society, it becomes more important to easily extract useful information from a large amount of texts. Question answering system is one of the information retrieval systems. It receives a question in natural language as an input, retrieves an answer to the question from a given document set or knowledge source, and returns it to a user.

General question answering systems consist of the following modules: question analysis, information retrieval, answer candidate extraction and answer candidate evaluation. First, in the question analysis module, morphological, syntactic and semantic analyses are performed to extract keywords used for information retrieval. In the information retrieval module, documents related to the keywords are retrieved. They are supposed to contain the answers of the question. In the answer candidate extraction module, several candidates of the answer are extracted from the retrieved documents. Finally, in the answer candidate evaluation module, scores of the answer candidates are calculated and show the answer with the highest score to the user.

Question answering systems are roughly classified into factoid and non-factoid types. Factoid question answering systems handle the questions to ask facts, things and names etc. Most of early research on the question answering systems focuses on factoid. On the other hand, non-factoid question answering systems handle the

questions asking more intelligent contents such as reason or a method. There are several types of non-factoid QA such as definition-type, why-type, how-type etc. Although factoid question answering systems are widely used as a kind of search engines, non-factoid QA systems are getting more important since they can retrieve cause of events or a way how to do something. It enables us to realize intelligent question answering.

The goal of this research is to carry out a survey on studies of inference scheme in why-question answering system and recognition of textual entailment that can handle a wide variety of linguistic expressions.

Why-question answering systems are based on retrieval of "causal statement" and "consequence statement" from the information source or documents. A causal statement is a text passage showing cause or reason of an event, while a consequence statement represents a consequence of the event. Morooka et al. proposed a rule-based why-question answering methods. A list of cue words indicating the cause or reason is constructed via a manual analysis on sample documents such as newspaper articles. Then patterns, which can extract a pair of a question and its answer or prevent an incorrect pair from being extracted from a document, are constructed. The patterns refer to the cue words and their surrounding words as well as their parts-of-speech (POSs). In addition, to extract not one but multiple sentences as the question or answer, a set of rules is built to extend a range of the question or answer when specific conjunctions appear at the boundary of them. They also made rules to remove redundant phrases from the answer according to an order of a contradictory conjunction and the answer. Finally, the similarity between the extracted question from the document and the question entered by the user is measured to evaluate validity of the answer candidate.

Another rule-based method guesses an order of a fact statement (sentence representing a fact or consequence) and a reason statement in a document. The rules are based on cue words, which are "reason word" (word indicating presence of a reason statement), anaphora and cataphora, in the fact statement. In answer candidate evaluation, a score of the pair of the fact and reason statements is calculated by weighted sum of frequency of interrogatives and reason words in the statements.

Since there are a wide variety of expressions of cause and reason, and much of them do not have explicit features, it is almost impossible to manually construct a set of rules to exhaustively find them. Higashinaka et al. attempt to automatically acquire

such rules by machine learning. Using the causal and non-causal sentences (sentences that express causal relation and not) in EDR corpus, linguistic patterns that frequently appear only in the causal sentences are acquired as the rules. Two kinds of the patterns are obtained. One is the pattern called ATS (Abstracted Text Span), which is a sequence of function words with wildcards that match any content words. The other is acquired by BACT (a Boosting Algorithm for Classification of Trees). It is similar to ATS pattern but POSs and semantic classes are also considered.

Harada et al. propose a method based on semantic network. The answer of the given question is extracted from the sentence in knowledge source that are the most similar to the question. The similarities of sentences are calculated by matching of semantic graphs of them. The semantic graph is a representation of a meaning of a sentence and obtained by semantic analysis identifying meaning of words and deep cases.

A method proposed by Tamura et al. generates keywords for search by paraphrasing the question and perform why-question answering by using existing Web search engines.

Syntactic structures can be used for re-ranking of the answer candidates in why-question answering. Verberne et al. report that important and useful syntactic structures are presence of a cue word, a main verb of the question, and relation between a focus of the question and a title of a document.

Tanaka et al. propose a method to train a classifier for identifying "why text segment", a portion of texts expressing cause or reason. Combination of base forms of function words and their POSs are called "Bag of Grammar" and used as features for training. This approach alleviates following problems: the learnt classifier is too specific to a domain; it is difficult to prepare labeled data for training.

Go et al. claim that positive/negative events tend to be caused by other positive/negative events (or reason), and propose a method to utilize semantic polarity of texts in why-question answering.

Cha et al. improve the performance of answer extraction by considering not only relation between the question and the answer candidate but also between the answer candidate and its surrounding sentences. Based on observation that a correct answer tends to appear with preceding/succeeding supplementary statement and be an important sentence in a document, the answer candidates are ranked by PageRank algorithm in a network where vertices are candidates.

Why-question answering systems are often required to extract the reason statements and fact statements from the document set. However, since the same content can be represented by various expressions, it is necessary to handle a variety of linguistic expressions. Research on diversity of texts, paraphrasing and implicature is called RTE (Recognizing Textual Entailment). It can be applicable for many natural language applications such as question answering system, summarization, machine translation and so on.

One of the methods to incorporate RTE module into a question answering system is to convert the question from interrogative to affirmative sentence, judge textual entailment between the affirmative sentence and the fact statement corresponding to the answer candidate, and re-rank the answer candidates based on reliability of RTE. There are two major approaches of RTE. One is investigation of algorithm of RTE, which focuses on similarity and alignment between two texts, transformation of a syntactic structure, formal logic, "Natural Logic" and so on. The other concerns how to develop useful knowledge for RTE.