Question Answering Handling Ambiguous Questions from Web

Wataru Osanai (0610019)

School of Information Science,
Japan Advanced Institute of Science and Technology

February 7, 2008

Keywords: Question Answering, Ambiguous Question, Web, Tables in Web Page.

This paper describes a concept of a open-domain ranking-type question answering system using Web. When a user inputs an ambiguous question, this system detects its ambiguity and outputs list of answers. Ambiguous questions in this research indicate ones that can't choose only one answer, because a meaning of word in user's question is ambiguous. For example, a question "world cup de yuushoushita kuni ha dokodesuka(Which country won the world cup?)" is ambiguous , because there are several "world cup" for soccer, ski, rugby and so on. When an user asks such a question, the system outputs list of answers with some additional information such as "Brazil(soccer world cup)", "Norway(ski world cup)" and "England(rugby world cup)". Previous research uses newspaper as knowledge resource. On the other hand, We use Web pages as knowledge resource. To extract list of answers like the above, we pay attention to table of web page. This is because list of answers may appear in a table in web pages. In this research, we describe the way to find an appropriate table from web page as the list of answers. Furthermore, the technique to generate a list of answers based on text analysis is also described.

The flow of the process of our system that outputs list of answers for ambiguous questions is as follows. First, keywords, answer type and keyword type are extracted from an user's question. There are two kinds of keywords, primary keyword and secondary keyword. Primary keyword is one that highly relates to the answer, which often appear as a topic of question. All other keywords are regarded as secondary keywords. Next, web pages are retrieved by a query of these keywords. If tables that include list of answers are successfully extracted from web page, those tables are presented. Otherwise, answer group is generated by text analyzing and presented.

The details of the way to extract tables that include list of answers are as follows. First, tables marked by table tags are extracted. Next, if the cell that exists in the first row or column in a table is equivalent to the primary keyword, these tables are extracted. Next, it is examined whether all secondary keywords exist in the (1)between title tags in the web page, (2)caption in the table or (3)three preceding segments from the table. When the secondary keyword doesn't exist, table is excluded from the candidate. Finally, it is examined whether the cells in the same row or column, which contains the primary keyword, include answer candidates. We calculate the ratio of number of cells such that Named Entity tag of a text in the cell is consistent with answer type to total number

of cells in the row or column. If the ratio is greater than 0.3, the table is extracted and presented to user.

The outline of the way to generate answer group by text analysis is as follow. First, retrieved web page is divided to the segments by HTML tag, and segments in which the answer seems to appear is retrieved. Next, answer candidates are extracted by NE tagging and syntactic pattern, etc. For keywords around extracted answer candidates, nouns highly related to it or nouns have a dependency relation with it, etc . are extracted as specializing expressions. "Specializing expression" is an expression that limits the meaning of ambiguous word, such as "soccer", "ski" and "rugby" in the previous example. Multiple triples (answer candidate, keyword, specializing expression) are extracted at this stage. The subsets of these triples where the keyword is common and specializing expression has some common attributes are get together as the answer groups. Because multiple answer groups are generally generated, answer groups are ranked by a score to choose the most appropriate one which stands for ambiguous meaning of a keyword. The score of answer group is defined according to (1)the number of specializing expression and answer candidate, (2)the type of common attributes of specializing expression, (3)score of answer candidates and (4)relevance between keywords and specializing expression, etc.

To evaluate our proposed method, 30 ambiguous questions were prepared. First, our system try to extract tables from web pages. When first step fails , the system try to generate answer group. According to the above procedure, we carried out experiments. As a result, tables from web pages or answer groups which are ranked best by scoring are the correct answer list for 56% of ambiguous questions. In addition, tables from web pages or answer groups which are ranked within 10th by scoring are the correct answer list for 83% of ambiguous questions. In these cases, for 9 of 30 questions, answer list were extracted by table extracting method, while for other questions, answer list were extracted by answer group generating. Furthermore, the number of question that can get correct answer list is increased by combining two techniques. The effectiveness of proposed method for using two techniques was confirmed by this experiment.