Exploitation of Newspaper-article Characteristics for Article Retrieval and Answer Extraction in QAC Task 2

Ruck THAWONMAS† Takayuki TOMOIKE††
Tomohiko KAWACHI††† Akio SAKAMOTO†††
† Department of Computer Science, Ritsumeikan University
1-1-1 Noji Higashi Kusatsu City, Shiga 525-8577, Japan
†† Course of Information Systems Engineering, Kochi University of Technology
††† Department of Information Systems Engineering, Kochi University of Technology
185 Miyanokuchi, Tosayamada-cho, Kami-gun, Kochi 782-8502, Japan
ruck@cs.ritsumei.ac.jp

Abstract

In this paper, we discuss a system for newspaper article retrieval and answer extraction. Due to the rapidly increasing amount of accessible information, systems that allow search in natural language are expected to play a much more important role in the very near future. Our system, called RAIK-Prassie, is designed for TASK 2 of QAC. The design of the RAIK-Prassie system focuses mainly on practical use of characteristics of newspaper articles and on questions related to person names. In addition, we implement the system such that it replies at most one answer to each query.
Keywords: *Newspaper Article Characteristics, Question Answering, Information Retrieval, Information Extraction*

1 Introduction

The amount of information spreading out over the Internet is dramatically increasing due to the popularization of WWW. Accordingly, WWW has become an indispensable tool to search/retrieve such information. However, compared with the quality and amount of transmitted information, still there are not sufficient tools and/or techniques for precisely and promptly searching/retrieving the designated information.

At present, the most commonly used information retrieval techniques are those by which users directly use keywords to specify the designated information. Though they impose fairly simple implementations to achieve information services, such techniques are hardly considered user-friendly. For example, sometimes, users may find it is difficult to specify a proper set of keywords, when a single keyword just does not work. In this case, the users are usually required to understand some search syntaxes such as AND/OR conditional statements. Thus, these techniques might not be considered as good candidates for precisely and promptly searching/retrieving the designated information.

To solve the above problem, researchers are currently developing information retrieval techniques that accept queries in natural language, cf., works in [1] and [2]. With these techniques, users can make queries to the systems as if they do to humans. For non-expert users, these techniques are hence much simpler to use compared to the keyword-based ones.

Question and Answering Challenge (QAC)[4], a part of the NTCIR Workshop, has been conducted which uses newspaper articles as benchmark data for information retrieval in natural language. We have participated in the TASK2 of QAC, and conducted research and development of our original system. Main features of our system include (a) exploitation of characteristics in newspaper articles and (b) specialization to queries related to person names.

In the previous version of our system, called RAIK-Prassie, [3], techniques for extraction of a single answer were proposed that use the distances among terms together with one characteristic in newspaper articles that important terms usually lie in the first sentence. Other newspaper article characteristics are further exploited in the current version of the RAIK-Prassie system, used in the Formal Run, discussed below in detail.

2 System Description

The RAIK-Prassie system is designed such that it will give at most one answer to a given query. In addition, it does not give an answer when the calculated reliability is below a pre-defined threshold.

The flowchart of the RAIK-Prassie system used in the Formal Run is shown in Fig. 1.

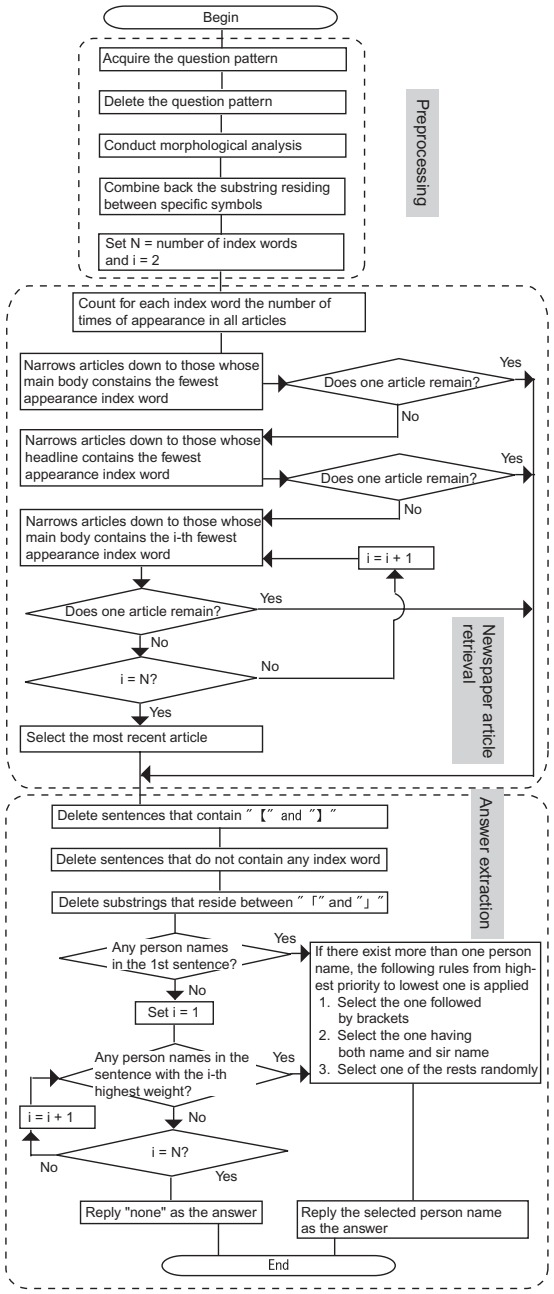

Figure 1: Flow chart of the RAIK-Prassie System.

Though the system is universally designed to cope with queries related to person names, places, times, quantities; this paper shall discuss mainly on how to deal with those queries related to person names.

2.1 Exploitation of Characteristics in Newspaper Articles

Newspaper articles must convey a relatively large amount of information within limited writing space. As a result, some relevant characteristics below can be seen.

- The main points of each article usually appear in the first sentence.
- Important matters are stated early in an article if not in the first sentence.
- Person names followed by brackets frequently indicate the highlighted persons in an article.
- The name of the reporter is given within the symbols "【" and "】".

The RAIK-Prassie system exploits the above characteristics to perform the QAC task 2.

2.2 Preprocessing

Given a query sentence, the RAIK-Prassie system first performs acquisition of the corresponding question pattern. For queries in Japanese, substrings implying such patterns normally come at the end of sentences. For example, in the query「ノーベル物理学賞を受賞した日本人は誰ですか。(Who are the Japanese laureates for the Nobel Prize for Physics?)」, the substring " は誰ですか "at the end of the sentence indicates that this query is asking for a person name. We call such a substring a question pattern. Acquisition of the corresponding question patterns allows us to specify the type of answers in the answer extraction part. A typical list of question patterns related to person names is shown in Table 1.

Table 1: A typical list of question patterns related to person names.

Pattern	Type
は誰でしたか	Who
は誰ですか	Who
は誰	Who

Once the answer type is known, the question pattern no longer contains important information. Hence, it is deleted from the query.

Next morphological analysis is done for the remaining query. Chasen [5] is used for this task. Chasen has been widely used as a tool for performing morphological analysis of Japanese sentences. This tool has also a function to describe the type of each segmented substring.

Usually substrings residing in the symbols such as "「" or "」" are also segmented after performing

morphological analysis. However, such substrings frequently represent proper nouns such as the titles of books, etc. They should therefore not be segmented. In our system, those terms are combined back to their original forms. The resulting segmented substrings are called index words.

2.3 Article Retrieval

Using the index words obtained in the preprocessing part, article retrieval is then performed that searches for articles containing those words. In the RAIK-Prassie system, the fewer frequently do the index words appear throughout the whole articles, the more important they are considered. Such index words are given higher weights.

In the article retrieval part, the following procedure is performed for narrowing down the article search space. The procedure stops when it finds one candidate article or does not find any.

1. Narrow down the articles to the ones whose main body contains the fewest appearance index word. If there are multiple resulting articles, go to the next step.
2. Narrow down the resulting articles to the ones whose headline contains the fewest appearance index word. If there are multiple resulting articles, go to the next step.
3. Iterate narrowing down the resulting articles to the ones whose main body contains the next fewest appearance index word. If there are multiple resulting articles, go to the next step.
4. Select the article whose published date is most recent.

2.4 Answer Extraction

From the resulting retrieved article, discussed below is how the answer is extracted.

First, the article is segmented. The following procedure is then executed.

- All sentences containing the following symbols ”「 ” and “ 」”are deleted. This is because such sentences have tendency to correspond to information related to the reporter and hence are not related to the answer of the query.
- All sentences that do not contain any index word are deleted. Such sentences are considered to represent auxiliary information in the article. Hence they have low tendency to contain the answer of the query.
- All substrings residing in the following symbols ”「” and ”」” are deleted. These particular symbols are usually used for dialogues or speeches of a highlighted person whose name, though, might be the answer.

Next, the answer is extracted. We employ a heuristic that the main points are stated in the first sentence. So if the first sentence contains a person name, we consider this person name as the answer. However, if there are multiple person names appearing in the first sentences, we select one of them according to the answer selection rules listed below from the highest priority to the lowest one.

1. Select the one followed by brackets.
2. Select the one having both name and sir name.
3. Select one of the rests randomly.

If the first sentence does not contain any person name, all other sentences are weighted according to the index terms they contain. Sentences with lower appearance index words are assigned higher amounts of weights. The sentence with the highest weight will be examined first to see whether or not it includes any person names. If a single person name exists in the sentence, it will be the answer. If multiple person names exist, then the above answer selection rules are applied. If this sentence contains no person name, then the sentence with the second highest weight will be examined. The procedure is repeated until a person name is eventually selected. If no person name exists in the article, the system outputs ”none” as its answer.

3 Formal Run Results

Results of eleven systems, participating in the QAC TASK 2, were announced by the QAC organizer. In QAC TASK 2, each system was asked to give answers to 200 given queries. These results are summarized in Table 2. In the table, the RAIK-Prassie system is given the identifier SysID S20005.

The definitions of the upper stack in a row are given as follows: Answer-the number of different answers in the task, Output - the number of answers that the user's system replied, Correct - the number of correct answers that the user's system replied. In addition, each value of the lower stack in a row is computed as follows:

- Recall = (the number of correct answers that the user's system output)/(the number of correct answers)

- Precision = (the number of correct answers that the user's system output)/(the number of answers that the user's system output)
- F-measure = (2 * Recall * Precision)/(Recall + Precision)

Table 2: Formal Run Results of Eleven Systems

SysID	Answer Recall	Output Precision	Correct F-measure
S20001	305 24.262	729 10.151	74 14.313
S20002	305 26.557	200 40.5	81 32.079
S20003	305 24.918	325 23.385	76 24.127
S20004	305 26.885	456 17.982	82 21.551
S20005	305 4.59	194 7.216	14 5.611
S20006	305 37.377	2000 5.7	114 9.892
S20007	305 12.131	276 13.406	37 12.737
S20008	305 20.656	2414 2.61	63 4.634
S20009	305 0	371 0	0 0
S20010	305 14.754	292 15.411	45 15.075
S20011	305 11.148	586 5.802	34 7.632
Ave.	305 18.480	713 12.924	56.364 13.423

We'd like to point out here that among 11 teams, though our system is ranked 10^{th} and 9^{th} for "Recall" and "F-measure", respectively, the system is ranked 7^{th} in terms of "Precision". This performance is fairly reasonable for a first time and relatively late participant team like ours. The RAIK-Prassie system has been designed to give only a single answer for a given query, though the query might essentially require a set of multiple answers. Another emphasis in the design is to have a system that will not give wrong answers to the users.

4 System Analysis

As shown in Fig. 1, the RAIK-Prassie system consists of three modules, namely, the preprocessing module, the newspaper article retrieval module, and the answer extraction module. In this section, we discuss our analysis results done for the newspaper article retrieval module. There are 200 queries in the Formal Run. Among them 43 queries are related to person names, which are our main targets for testing the developed system. In our analysis below, we focus on these 43 queries.

There are 18 queries for which wrong articles were retrieved. By wrong articles, we mean articles that do not contain the correct answers. Among these, 13(72%) articles were retrieved because they contain the fewest appearance index word in their headline. However, for the remaining 25 queries, the RAIK-Prassie system could retrieve articles that contain the correct answers for 22(88%) queries because those articles contain the fewest appearance index word in their headline.

As a result, our heuristic to retrieve articles according to what their deadline contains might not be as viable as we thought. Reconsideration of the procedure for narrowing down articles is necessary.

5 Remaining Problems and Future Work

The Formal Run version of the RAIK-Prassie system retrieves only one article and attempts to extract an answer from the retrieved article. However, as discussed above, we have found that retrieved articles might not always contain the correct answers (Table 3). To solve this problem, we are now extending the article retrieval module such that it allows retrieval of multiple articles by combining the narrowing down results from both article bodies and headlines.

Table 3: Successful and Failed Examples

Successful Example	
Query ID	QAC1-2033-01
Query	「速水優の前の日銀総裁は誰ですか。 (Who was the head of the Bank of Japan before Yu Hayami?)」
Article ID	980414353
Answer	「松下康雄 (Yasuo Matsushita)」
Failed Example	
Query ID	QAC1-2063-01
Query	「源頼朝の弟は誰ですか。 (Who is the younger brother of the shogun Yoritomo Minamoto?)」
Article ID	991228141
Answer	none (could not find the answer)

In addition, the length of each newspaper article in the repository varies. As a result, in the article retrieval module, articles with longer lengths have higher tendency to be retrieved than those with shorter lengths. This is a problem because longer articles do not always contain correct answers. To solve this, we are applying text abstraction techniques such as Posum[6] which is publicly available. The resulting abstract of each article should have similar length. This might improve the performance of our system.

References

[1] Wataru HIGASA. Dialogue Helpsystem based on Flexible Matching of User Query with Knowledge Base. Master Thesis, Kyoto University, 2000 (in Japanese).

[2] Takayuki TOMOIKE. A Study on Construction of a Log Analysis System Utilizing Similar Case Data. Senior Thesis, Kochi University of Technology, 2001 (in Japanese).

[3] Takayuki TOMOIKE, Ruck THAWONMAS, and Akio SAKAMOTO. A Consideration on Information Retrieval and Information Extraction in a Question Answering System. Proc. Shikoku Information Processing of Japan Research Symposium, 2002 (in Japanese).

[4] QAC Homepage: http://www.nlp.cs.ritsumei.ac.jp/qac/

[5] Yuji MATSUMOTO. Morphological Analysis System ChaSen version 2.2.8 Manual. NAIST, Matsumoto Laboratory, 2001.

[6] Hajime MOCHIZUKI. Posum version 1.50.2 Manual. JAIST, 2002.