Proceedings of NTCIR-4, Tokyo, April 2003 - June 2004

Question Answering System QUARK

Yasuo Nii Keizo Kawata Tatsumi Yoshida Hiroyuki Sakai Shigeru Masuyama
Department of Knowledge-based Information Engineering, Toyohashi University of Technology
1-1 Hibarigaoka, Tempaku-cho, Toyohashi, Aichi 441-8580, Japan
masuyama@tutkie.tut.ac.jp

Abstract

Recently, we can acquire immence amount of information thanks to the spread of a computer and internet. Therefore, technology for finding the information that a user desires becomes more and more important. A question answering (QA) system answers a question written by natural language in contrast to conventional information retrieval systems where a user expresses his information need by keywords. Therefore, a QA system can provide more user-friendly information access environment to a user. We developed a QA system QUARK[1] *that finds answer words from large newspaper article corpora and was evaluated by participating in NTCIR4 QAC2, an evaluation workshop of Japanese QA sytems.*

Keywords:
question and answering system, heuristic rules for answer classification, hypernym acquisition from a corpus for classification of the answer candidate words

1 Introduction

Recent rapid advance of a computer and internet has enabled us to acquire immence amount of information. To support a user to access appropreate information from such vast sea of information, information retrieval technology has become more and more important. Search engines like Google accept keywords as an input by assuming that these keywords represent the user's information need. Then, a user receives the list of documents that contains keywords. However, it is difficult to express user's information need precisely by using keywords and, the information that a user wishes to obtain is not always included in a document among the document lists presented by the system. Therefore, the user must manually select only an actually necessary document from the retrieved documents.

[1]QUestion AnsweRing system using a large corpus as a Knowledger source

A question answering (QA) system answers a question written by a natural language in contrast to conventional information retrieval systems where a user expresses his information need by keywords. For example, when a demand for information "I want to know the first Japanese Prime Minister." occurs, the answer "Ito Hirofumi" is presented on the display by inputting the question sentence described in the natural language, e.g., "Who is the first prime minister of Japan?". Therefore, a QA system can provide more user-friendly information access environment to a user than conventional information retrieval systems.

We develop a QA system QUARK that finds answer words from large newspaper article corpora in this project. QUARK was evaluated by participating in NTCIR4 QAC2, an evaluation workshop of Japanese QA sytems. Detailed information on this workshop is found in (Fukumoto, et al., 2004)(Kato et al., 2004).

QUARK has the following two features.

Firstly, when expected answer classification such as a person's name and a company name, the detailed answer classification that can not be obtained only with an interrogative is estimated by using manually-made rules. Moreover, when answer classification is a numeral, easy-to-take unit expressions are also estimated.

Secondly, the hypernym mentioned with a common noun is acquired from a corpus to the unknown answer candidate word in the thesaurus, and the classification of the answer candidate word is estimated based on that hypernym. This enables QUARK to compare answer candidates not in the thesaurus like a named entity with the answer classification acquired from a question sentence. The comparison of the performance among systems is done objectively with QAC2 of NTCIR4 by giving them the same question respectively.

Problems given to a QA system at QAC2 are limited to those whose answer is mentioned in a source of knowledge, and those can be answered with a word, and excluding those requiring reasoning, e.g., "*naze*" or "*nani*".

2 System Overview of QUARK

QUARK consists of two modules, the document retrieval part and the answer extraction part, respectively.

The article that seems to contain an answer is searched by the document retrieval part by using key words contained in the input question sentence from the knowledge source.

On the other hand, an actual answer word is extracted in the answer extraction part from the document set obtained by the document retrieval part. First, only the answer candidate word corresponding to the question form after estimating answer expected by the question sentence. Finally, assign a score to each answer candidate considering the position the keyword contained in the question sentence etc., and the answer candidate word having the high rank is presented to every answer candidate to the user as an answer.

3 Document retrieval part

QUARK retrieves documents that include the answer from a newspaper article corpora. The outline of the procedure is shown below.

1. Let words contained in a question sentence be keywords and each keyword is assigned weight w_i. Note that words in the predetermined stop word table are excluded.

2. Article A_i's score $score(A_i)$ is calculated by the keyword included in the article.

3. Articles with the $score(A_i)$ below predetermend threshold value are removed.

Details of 1 above is as follows: First, a morphological analysis is applied to the question sentence given to it, and a set of morphemes is obtained. We used JUMAN, Ver.4.0 as a morphological analyzer. Then, by using manually constructed stop word table, words such as the interrogative unnecessary for the retrieval, and a form noun are removed. Only nouns, adjectives and adverbs are left in the set Q of the retrieval words. Note that a compound word is regarded as a word.

The given document set is retrieved using a query of the form of

$\bigvee_{q \in Q} q$

For the set A of retrieved documents by the query, we assign, in 2 above, score to $a_i \in A$ by the following formula:

$$score(a_i) = \sum_{k \in K(a_i, Q)} (w(k) + \log tf(a_i, k))$$

Here, $K(a_i, Q)$ is the set of keywords in article a_i, $w(k)$ is the weight of a keywords k, and $tf(a_i, k)$ is the number of keywords in a_i.

Weight of a keyword is determined, by a preliminary experiments, as follows:

- The clause containing keywords is parenthesized by "「" and "」" (parentheses commonly used in Japanese sentences): $+30$
- Keyword k is a numerical expression: $+15$
- Keyword k is a location name: $+10$
- Keyword k is a person's name: $+10$
- Keyword k is a compound word: $+5$
- Keyword k is a "*katakana*" word: $+5$
- Keyword k is contained in the first clause in the question sentence: $+10$
- Otherwise: 10

Note that classification of words are done by considering the output of KNP.

4 Answer extraction part

In the answer extraction part, after a question form that represents an answer expected by the question sentence is estimated, a question sentence leaves only the answer candidate word corresponding to the question form.

Finally, a score is assigned to each answer candidate word in consideration of the one related to the position with the keyword contained in the question sentence and so on, and the answer candidate word having high rank is presented to the user as an answer.

The nouns collocating in the same sentence with the keyword used in the document retrieval part are used as preliminary answer candidate words.

4.1 Estimation of the Question Type

The question form is an answer classification expected by an input question sentence as an answer. For example, "person's name" becomes a question form in the question "Who is the author of *Botchan*", and "year|month|date" becomes a question form in the question "*When will the fifth edition of koujien(a Japanese dictionary) be sold?*" Here | denotes or. Note that the question form actually assigned by QUARK is the classified name of the thesaurus made by Sekine et al.[2]

Furthermore, a unit expression is assigned when a question form is a numeral. Therefore, it is assigned in the following form.

Example 1 "Who is the author of *Botchan*?" (person)

Example 2 "When will the fifth edition of koujien be sold?" (date |month | year)

Whether a question form is a noun or a numeral is decided by the kind of interrogatives being used in the question sentence, which are shown below:

nani (what), *doko*(where), *dare* (who), etc.: nouns

[2] A hierarchical thesaurus released only for participants for NTCIR4 QAC2. 78,017 named entities and 15,843 common nouns are classified into about 200 classes.

itsu (when), *ikura*(how much), *nani*+unit(meter, etc.): numerals

4.2 When the question form is a noun

When a question form is a noun, it may be estimated only with an interrogative like "*... ha dare-desuka?*", or it may not be estimated only with an interrogative like "*... ha nan desuka*". Thus, a question form is estimated by QUARK by referring to the decision word contained in case of the latter, while a question is asked to it from the interrogative in case of the former.

The decision word denotes the word contained in the question sentence and whose thesaurus classification becomes a question form.

For example, in the question "*Who wrote Botchan?*", "novel" becomes a decision word, and a decision word itself and classification on that thesaurus "novel, book" becomes a question form. Moreover, the classification of the decision word due to the thesaurus is decided uses the same method as that of classifying the answer candidate word shown in the next section.

The example of the rule to distinguish a decision word is shown as follows.

Example of question matched to the rule and the question type "*dare*": *...ha dare ga okonaimasitaka* →person

"noun A *ha nani*": *...ga kaita syousetsu ha nani* → work

"noun A *no namae ha nani*": *...wo hatsubaishita kaisha nonamae ha nani* → company

4.3 When the question form is a numeral

A question form for a numeral can be determined easily when an interrogative where corresponding unit expression can be decided uniquely is used like example 3 or when the unit expression is specified like example 4.

Example 3 "When did "Matsumoto Salin" matter happen?" (date | month | year)

Example 4 "What is the population of Japan?"(*nin*)

However, in such a case, "an interrogative *nani* + unit expression" is not always used. For example, like Example 5, the case when a unit expression is not described to the question sentence directly often happens.

Example 5 "How much is the price of PS2?"(yen | dollar etc.)

Therefore, when the unit expression that a question sentence is asked for as an answer is not described directly, and the estimation of the unit expression that the answer word is easy to be included is done by a statistical method. Outline of the estimation of question form when question form is numeral is as folllows:

Table 1. Matrix for calculating χ^2 value

	N_3	N_4	total
N_1	a	b	e
N_2	c	d	f
total	g	h	n

1. Among phrases modifying interrogative expression let those having particle *ha* removing particles be keyword 1, and let phrases modifying as "*no* case words in keyword 1 removing particles be keyword 2. (keyword 1 is "price" and keyword 2 is PS2 in the case of Example 5.)

2. Search sentences containing both keywords 1 and 2.

3. Judge the independence of the keyword 1 and each unit expression (obtained from NTT *Goitaikei*(a Japanese thesaurus)) (Ikehara et al., eds., 1997) by χ^2 test extracted from the collected sentence set.

4. Regard those unit expression judged to be not independent be the unit expression that the answer to the question has.

When independence with the keyword 1 and the unit expression is judged by χ^2 test, 2×2 matrix shown in table 1 is made and calculated by using the following formula.

$$\chi^2 = \frac{n(|ab-bc|-n/2)^2}{efgh}$$

Here, in the matrix shown in Table 1, N_1 is the number of sentences containing unit A, N_2 is the number of sentences not containing unit A, N_3 is the number of sentences containing keyword 1, N_4 is the number of sentences not containing keyword 1, a is defined as the number of times when the phrase which contained a unit A is modified by keyword 1 directly or when the phrase which a keyword 1 contained a unit A to in the same sentence directly. In other words, a is counted when a sentence such as examples 6 and 7 exist when keyword 1 is length, and unit A is a meter.

Example 6 "The length of the object is 1 meter. "

Example 7 "An object with length of 10 meters."

4.4 Estimation of Answer Extraction Word Type

QUARK compares answer extraction word type with the question type. It keeps only the same type word. But in many named entities, it is impossible for the system to refer its classification from a dictionary. Therefore, QUARK uses a method that extracts a set of named entities and hypernyms (common nouns) from a corpus, and a classification of hypernyms is used for classification of named entities.

The method of extracting hypernym uses the two features of newspaper articles. The first feature is that

a named entity often appears together with its hypernym in an article. The second feature is that they have patterns in the way of expression. For example, 'named entity' *dearu* 'hypernym' (e.g., "*sekai saidai no 'hana' dearu* 'rafureshia'(The biggest flower in the world Rafflesia)).

On the other hand, this method is not effective to person class (for example, "*yamada shachou*"(president Yamada)), because they appear in the same segment. Therefore, QUARK uses the classification technique using SVM to classify the person class.

4.4.1 When the answer candidate is a common noun

When an answer candidate word is a common noun, it is classified using the previously mentioned thesaurus by Sekine et al. But, as for the case where a common noun is a compound word, because only a few compound words are described in the thesaurus directly, classification is done by the following four rules. Rule 1 has the highest and rule 4 has the lowest priority, respectively, in applying the rules.

Rule 1 When the compound word is in the thesaurus, the classification by the thesaurus is adopted.

Rule 2 Among the suffices included in a compound word, those in the thesaurus and has the longest length is selected. The classification which that suffix belongs to is made the classification of the compound word.

Rule 3 Among the prefix contained in the compound word, the longest one in the thesaurus is chosen. The classification of the suffix is regarded as the classification of the compound word.

Rule 4 The classification that the word having the longest common suffix among words having a common suffix in the thesaurus is made the classification of the compound word.

4.4.2 When the answer candidate is a named entity

When an answer word candidate is a named entity, it is seldom in the dictionary, and, therefore, it is more difficult to acquire a class directly from the thesaurus for a named entity than word's being in the dictionary. Hence, the hypernym of the named entity is acquired from the corpus, and this problem is tried to be solved by QUARK by making the classification of the hypernym be the classification of the named entity.

As for the advantage of the acquisition of the hypernym from the corpus is that information lack due to generarization of the named entity is lower comparing with the case when a dictionary is used. In particular, when a corpus consists of newspaper articles, a hypernym with more amounts of information is acquired than those of the classification acquired from the thesaurus. Because of that, more powerful focusing can be attained in the choice of the answer candidate word. It is also easy to generalize the hypernym acquired from corpus by using the technique to ask the classification of the common noun previously introduced for the classification using the thesaurus if necessary. The characteristics of the description of the newspaper, "When a named entity a reader can not judge appears, its hypernym is often used for its explanation.", are used for the acquisition of the hypernym as illustrated in examples 8 to 10.

Example 8 A major communication company "NTT"

Example 9 The biggest flower in the world Rafflesia.

Example 10 An automobile called Corolla.

A template is used when a hypernym is actually acquired. Some patterns can be found by the newspaper articles in the method of the indication of the named entity and the hypernym. For example, patterns e.g., "named entity 「hypernym」 ", "hypernym *dearu* named entity", "named entity *to yobareru* hypernym" can be considered for the above examples.

By adopting patterns whose appearance frequency is high as templates, QUARK attempts to acquire the hypernym of the unknown named entity by templates. How to collect templates is shown in the following.

1. Pair of nouns A and B is extracted by using the template of "noun A 「noun B」 " from the corpus by pattern matching.

2. The sentences where noun B collocates with noun A are collected with an expression except for "noun A 「noun B」 ". When noun B appears after the second time, the explanation is often omitted, thus only the sentence that noun B appears for the first time is considered.

3. The partial string put between nouns A and B from the collected sentence is extracted. Then, by substituting the part where noun A is located with the hypernym and the part where noun B is located with the named entity.

4. Among acquired templates, those with frequency under the predetermined threshold are removed.

4.5 Scoring to Answer Candidates

The score of an answer candidate is calculated by the following formula.

$$W(w_i) = (\frac{sf(w_i)}{S})\log(\frac{N}{df(w_i)}) \times \frac{\sum_{k_j \in query} dist(k_j, w_i)}{length(s_{w_i})}$$

Let S be the number of sentences contained in articles retrieved at the document retrieval part, N be the number of sentences contained in corpora, $sf(w)$ be the number of sentences contaning word w_i in corpora, $dist(k_i, w_i)$ be distance k_i and w_i in the sentence containing word w_i, $length(s_{w_i})$ be the length of sentence

Table 2. Rate of correct answers for each question form

Question form	Q	C	R
person's name	48	26	54.1
artifact	37	16	43.2
natural things	14	8	57.1
location	31	14	45.2
organization	23	13	56.5
numerals	23	12	52.2
date, month or year	9	6	66.7
others	14	4	28.6
total	200	99	49.5

s_{wi}, and a query be the set of keywords contained in the given question.

5 Experiments for evaluation

The above technique was implemented and QUARK was evaluated by participating in QAC2 of NTCIR4.

Experiments are done under the following conditions;

Problem: 200 questions for Task 1 of QAC 2.

Knowledge Source: Mainichi Shimbin, 1998, 1999. Yomiuri Shimbun, 1998, 1999.

External Knowledge: Thesaurus made by Sekine et al, NTT Goitaikei(a Japanese thesaurus).

Tools:

- search engine Namazu Ver.2.0.12
- morphological analyzer JUMAN ver.4.0
- parser KNP Ver.2.0b6

Evaluation Criteria: That of Task 1 of QAC 2

Results are shown as follows:

The number of questions is 200, the number of answers is 392, the number of answers by the system is 996, the number of correct answers is 122, respectively. Moreover, recall is 0.311, precision is 0.122, and MMR is 0.344, respectively. MMR of QUARK is the 15th best among 25 systems participated in Task 1 of QAC2.

Table 2 shows the rate of correct answers for each question form. Here Q is the number of questions, C is the number of correct answers, and R is the rate of correct answers.

6 Discussion

The number of correct answers was 122, and the number in question which contains a correct answer was 99 questions with a result of QAC2 Task1. As for the correct answer rate of every question form, QUARK could not necessariry get a high correct answer rate although a person's name, a nature name, an organization name, a numeral and a date, month and year were beyond 50%.

In most cases an answer was not contained in the article retrieved in the document retrieval part when causes of the errors are examined, which shares about 50% of total errors. Errors caused at the scoring stage and the classification errors of answer candidates follow.

6.1 Errors at the document retrieval

The recall and precision of the document retrieval part are 0.32 and 0.30, respectively. The recall shows the ratio of successful extraction of articles containing the correct answers, while precision shows the ratio of articles containing the correct answers among those retrieved. In particular, low precision is not the direct cause of errors to show but it influences the score of the answer because the number of the unnecessary answer words increases due to low precision.

Difference between keyword for the article retrieval and the notation in a source of knowledge causes decrease of the recall rate. The decrease of recall rate was due to deciding the importance of the keyword statically.

In QUARK, the importance of the keyword contained in the question sentence is decided only by its kind and it does not use either the parsing information of the question sentence or that of the question form. Thus, the unnecessary keyword contained in the question sentence can not be omitted, and the article that contains many unnecessary articles is extracted.

6.2 Errors at the score calculation

An equation considering $tf \bullet idf$ and relative position with the keyword contained in the question sentence is used in this system. However, the frequency of the answer words considerablly affects the equation used in this system. Hence, an answer word itself must appear more than one article, and its frequency must be large enough to increase the value of the score. Therefore, we may consider that it causes decrease of correct answer rate of a question having answer words whose frequency is small.

6.3 Errors at the classification of answer candidate words

The classification of the answer candidate word is done by extracting a hypernym from a corpus. However, as for the named entity, it exists abundantly when that hypernym is contained in the named entity itself

like an example 11. In such a case, which caused errors of the answer candidate word cannot be coped with by our technique to acquire a named entity and an hypernym in the phrase unit.

Example 11: Artificial satellite Himawari, University of Tokyo, Director Kurosawa.

Moreover, as for the work of art such as a movie and a play, a hypernym was not mentioned directly, and there was an example, e.g., Example 12, where the word can be judged as a work of art from the context.

Example 12: Carmen was presented in Aoyama theater.

6.4 Errors at the classification of question form

Estimation in question form uses manually-made rules. However, because rules were manually made, they have not large coverage, and there were fourteen examples that classification was mistaken. Moreover, there were thirty examples to which no rule is applicable besides direct errors. In this case, an answer is decided to be extracted only using the score, and it causes more decrease of the precision.

6.5 Other errors

There are a compound errors of the same answer as others and a cause of the precision decreases. The erroneous answer of the plural which has the same meaning occupies a high rank.

7 Task 3

We briefly introduced our preliminary approach to Task 3 of QAC3, For its detailed task description, see (Kato et al., 2004). QUARK come the 10th best position among 14 participating systems.

7.1 Method

As, in many cases, related question may not contain information obtainable from the original question, information from the related question is insufficient for obtaining the correct answer. Only a few keywords may be obtained from a related question, and the article group obtained in the document retrieval part becomes huge in quantity. To cope with these problems, one characteristic word is chosen from keywords of the original question, and add to the keyword of the related question in the document retrieval part. The proposed procedure is shown below.

1. Delete the word not suitable as an additional word.

The word corresponding to the conditions shown below is deleted from keywords of the question.

- The word that overlaps keyword of the related question.
- Numeral accompanied with a unit expression and the same unit expression exists in the related question.

2. Each keyword remained is given the following weight, where a modified tf•idf method is used.

$$P(W_i) = (\frac{tsf(W_i)}{S}) \log(\frac{N}{df(W_i)}), \qquad (1)$$

where $P(W_i)$ is the weight of keyword W_i, $W_i,\ i = 1, 2, \ldots, n$ are keywords, N is the number of all documents in a corpus, $df(W_i)$ is the number of documents containing the keyword W_i, S is the number of articles obtained by document retrieval at the time of analysis of the original question, and $tsf(W_i)$ is the number of the documents containing the keyword W_i.

3. A keyword with the largest weight is added to keywords of related question.

8 Conclusion

We developed a question answering system QUARK that outputs the words and phrases as an answer to the question sentence input written in the natural language. Then, we participated in the evaluation workshop QAC2 for question answering. With evaluation in Task1 of QAC2, 49% of 99 questions were correctly answered among 200 questions. Moreover, among the 5th best answers output by the system, the number of correct answers existed in the 1st best was 49 questions of the 22%. Recall and precision attained 31.1% and 12.2%, respectively.

Acknowledgement This work was supported in part by The 21st Century COE Program "Intelligent Human Sensing", from the Ministry of Education, Culture, Sports, Science and Technology of Japan, and The Grant-in-Aid ((C)(2)13680444) from the Japan Society for the Promotion of Science.

References

[1] Jun'ichi Fukumoto, Tsuneaki Kato and Fumio Masui. 2004. Question answering challenge for five ranked answers and list answers–an overview of NTCIR4 QAC2 Subtask 1 and 2–, to appear in Working Notes NTCIR4.

[2] Satoru Ikehara, et al., eds. 1997. NTT *Goitaikei*(a Japanese thesaurus), Iwanami Publishing.

[3] Tsuneaki Kato, Jun'ichi Fukumoto and Fumio Masui. 2004. Question answering challenge for information access dialogue–overview of NTCIR4 QAC2 subtask 3–, to appear in Working Notes NTCIR4.