Non-Factoid Japanese Question Answering through Passage Retrieval that Is Weighted Based on Types of Answers

Masaki Murata and Sachiyo Tsukawaki

National Institute of Information and
Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun,
Kyoto 619-0289, Japan
{murata,tsuka}@nict.go.jp

Toshiyuki Kanamaru

Kyoto University
Yoshida-Nihonmatsu-Cho, Sakyo
Kyoto, 606-8501 Japan
kanamaru@hi.h.kyoto-u.ac.jp

Qing Ma

Ryukoku University
Otsu, Shiga, 520-2194, Japan
qma@math.ryukoku.ac.jp

Hitoshi Isahara

National Institute of Information and
Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun,
Kyoto 619-0289, Japan
isahara@nict.go.jp

Abstract

We constructed a system for answering non-factoid Japanese questions. We used various methods of passage retrieval for the system. We extracted paragraphs based on terms from an input question and output them as the preferred answers. We classified the non-factoid questions into six categories. We used a particular method for each category. For example, we increased the scores of paragraphs including the word "reason" for questions including the word "why." We participated at NTCIR-6 QAC-4, where our system obtained the most correct answers out of all the eight participating teams. The rate of accuracy was 0.77, which indicates that our methods were effective.

1 Introduction

A question-answering system is an application designed to produce the correct answer to a question given as input. For example, when "What is the capital of Japan?" is given as input, a question-answering system may retrieve text containing sentences like "Tokyo is Japan's capital and the country's largest and most important city", and "Tokyo is also one of Japan's 47 prefectures", from Websites, newspaper articles, or encyclopedias. The system then outputs "Tokyo" as the correct answer. We believe question-answering systems will become a more convenient alternative to other systems designed for information retrieval and a basic component of future artificial intelligence systems. Numerous researchers have recently been attracted to this important topic. These researchers have produced many interesting studies on question-answering systems (Kupiec, 1993; Ittycheriah et al., 2001; Clarke et al., 2001; Dumis et al., 2002; Magnini et al., 2002; Moldovan et al., 2003). Evaluation conferences and contests on question-answering systems have also been held. In particular, the U.S.A. has held the Text REtrieval Conferences (TREC) (TREC-10 committee, 2001), and Japan has hosted the Question-Answering Challenges (QAC) (National Institute of Informatics, 2002) at NTCIR (NII Test Collection for IR Systems) 3. These conferences and contests have aimed at improving question-answering systems. The researchers who participate in these create question-answering systems that they then use to answer the same questions, and each system's performance is then evaluated to yield possible improvements.

We addressed non-factoid question answering in NTCIR-6 QAC-4. For example, when the question was "Why are people opposed to the Private Information Protection Law?" the system retrieved sentences based on terms appearing in the question and output an answer using the retrieved sentences. Numerous studies have addressed issues that are involved in the answering of non-factoid questions (Berger et al., 2000; Blair-Goldensohn et al., 2003;

727

Xu et al., 2003; Soricut and Brill, 2004; Han et al., 2005; Morooka and Fukumoto, 2006; Maehara et al., 2006; Asada, 2006).

We constructed a system for answering non-factoid Japanese questions for QAC-4. We used methods of passage retrieval for the system. We extracted paragraphs based on terms from an input question and output them as the preferred answers. We classified the non-factoid questions into six categories. We used a particular method for each category. For example, we increased the scores of paragraphs including the word "reason" for questions including the word "why." We performed experiments using the NTCIR-6 QAC-4 data collection and tested the effectiveness of our methods.

2　Categories of Non-Factoid Questions

We used six categories of non-factoid questions in this study. We constructed the categories by consulting the dry run data in QAC-4.

1. Definition-oriented questions (Questions that require a definition to be given in response.)

   e.g., *K-1 to wa nandesuka?* (What is K-1?)

2. Reason-oriented questions (Questions that require a reason to be given in response.)

   e.g., *kojin jouhou hokogou ni hantai shiteiru hito wa doushite hantai shiteiru no desuka?* (Why are people opposed to the Private Information Protection Law?)

3. Method-oriented questions (Questions that require an explanation of a method to be given in response.)

   e.g., *sekai isan wa donoyouni shite kimeru no desuka?*" (How is a World Heritage Site determined?)

4. Degree-oriented questions (Questions that require an explanation of the degree of something to be given in response.)

5. Change-oriented questions (Questions that require a description of things that change to be given in response.)

   e.g., *shounen hou wa dou kawari mashitaka?* (How was the juvenile law changed?)

6. Detail-oriented questions (Questions that require a description of the particulars or details surrounding a sequence of events to be given in response.)

   e.g., *donoyouna keii de ryuukyuu oukoku wa nihon no ichibu ni natta no desuka?* (How did Ryukyu come to belong to Japan?)

3　Question-answering Systems in this Study

The system has three basic components:

1. Prediction of type of answer

   The system predicts the answer to be a particular type of expression based on whether the input question is indicated by an interrogative pronoun, an adjective, or an adverb. For example, if the input question is "Why are people opposed to the Private Information Protection Law?", the word "why" suggests that the answer will be an expression that describes a reason.

2. Document retrieval

   The system extracts terms from the input question and retrieves documents by using these terms. Documents that are likely to contain the correct answer are thus gathered during the retrieval process. For example, for the input question "Why are people opposed to the Private Information Protection Law?", the system extracts "people," "opposed," "Private," "Information," "Protection," and "Law" as terms and retrieves the appropriate documents based on these.

3. Answer detection

   The system separates the retrieved documents into paragraphs and retrieves those that contain terms from the input question and a clue expression (e.g., "to wa" (copula sentence) for the definition sentence). The system outputs the retrieved paragraphs as the preferred answer.

3.1　Prediction of type of answer

We used the following rules for predicting the type of answer. We constructed the rules by consulting the dry run data in QAC-4.

1. Definition-oriented questions Questions including expressions such as "*to wa nani,*" "*donna,*" "*douiu,*" "*douitta,*" "*nanimono,*" "*donoyouna mono,*" "*donna mono,*" and "*douiu koto*" (which all mean "what is") are recognized by the system as being definition-oriented questions.

2. Reason-oriented questions Questions including expressions such as "*naze*" (why), "*naniyue*" (why), "*doushite*" (why), "*nani ga riyuu de*" (what is the reason), and "*donna riyuu de*" (what reason), are recognized by the system as being reason-oriented questions.

3. Method-oriented questions Questions including expressions such as "*dou,*" "*dousureba,*" "*douyatte,*" "*dono youni shite,*" "*ikani shite,*" "*ikani,*" and "*donna houhou de*" (which all mean "how") are recognized by the system as being method-oriented questions.

4. Degree-oriented questions Questions including expressions such as "*dorekurai*" (how much), "*dorekurai no*" (to what extent), and "*dono teido*" (to what extent), are recognized by the system as being degree-oriented questions.

5. Change-oriented questions Questions including expressions such as "*naniga chigau*" (What is different), "*donoyuni kawaru*" (How is ... changed), and "*dokoga kotonaru*" (What is different), are recognized by the system as being change-oriented questions.

6. Detail-oriented questions Questions including expressions such as "*dono you na keii,*" "*dono you na ikisatsu,*" and "*dono you na nariyuki*" (which all mean "how was") are recognized by the system as being detail-oriented questions.

3.2 Document retrieval

Our system extracts terms from a question by using the morphological analyzer, ChaSen (Matsumoto et al., 1999). The analyzer first eliminates prepositions, articles, and similar parts of speech. It then retrieves documents by using the extracted terms.

The documents are retrieved as follows:

We first retrieve the top k_{dr1} documents with the highest scores calculated using the equation

$$Score(d)$$
$$= \sum_{\text{term } t} \left(\frac{tf(d,t)}{tf(d,t) + k_t \frac{length(d) + k_+}{\Delta + k_+}} \times log \frac{N}{df(t)} \right),$$
(1)

where d is a document, t is a term extracted from a question, and $tf(d,t)$ is the frequency of t occurring in d. Here, $df(t)$ is the number of documents in which t appears, N is the total number of documents, $length(d)$ is the length of d, and Δ is the average length of all documents. Constants k_t and k_+ are defined based on experimental results. We based this equation on Robertson's equation (Robertson and Walker, 1994; Robertson et al., 1994). This approach is very effective, and we have used it extensively for information retrieval (Murata et al., 2000; Murata et al., 2001; Murata et al., 2002). The question-answering system uses a large number for k_t.

We extracted the top 300 documents and used them in the next procedure.

3.3 Answer detection

In detecting answers, our system first generates candidate expressions for them from the extracted documents. We use two methods for extracting candidate expressions. Method 1 uses a paragraph as a candidate expression. Method 2 uses a paragraph, two continuous paragraphs, or three continuous paragraphs as candidate expressions.

We award each candidate expression the following score.

$$Score(d)$$
$$= -min_{t1 \in T} log \prod_{t2 \in T3} (2dist(t1,t2)\frac{df(t2)}{N})$$
$$+ \quad 0.00000001 \times length(d)$$
$$= \quad max_{t1 \in T} \sum_{t2 \in T3} log \frac{N}{2dist(t1,t2) * df(t2)}$$
$$+ \quad 0.00000001 \times length(d)$$
(2)

$$T3 = \{t | t \in T, 2dist(t1,t)\frac{df(t)}{N} \leq 1\}, \qquad (3)$$

where d is a candidate expression, T is the set of terms in the question, $dist(t1,t2)$ is the distance between $t1$ and $t2$ (defined as the number of characters between them with $dist(t1,t2) = 0.5$ when $t1 = t2$), and $length(d)$ is the number of characters in a candidate expression. The numerical term, $0.00000001 \times length(d)$, is used for increasing the scores of long paragraphs.

For reason-oriented questions, our system uses some reason terms such as "riyuu" (reason), "gen'in" (cause), and "nazenara" (because) as terms for Eq. 2 in addition to terms from the input question. This is because we would like to increase the score of a document that includes reason terms for reason-oriented questions.

For method-oriented questions, our system uses some method terms such as "houhou" (method), "tejun" (procedure), and "kotoniyori" (by doing) as terms for second document retrieval (re-ranking) in addition to terms from the input question.

For detail-oriented questions, our system uses some method terms such as "keii" (a detail, or a sequence of events), "haikei" (background), and "rekishi" (history) as terms for second document retrieval (re-ranking) in addition to terms from the input question.

For degree-oriented questions, when candidate paragraphs include numerical expressions, the score ($Score(d)$) is multiplied by 1.1.

For definition-oriented questions, the system first extracts focus expressions. When the question includes expressions such as "*X-wa*", "*X-towa*", "*X-toiunowa*", and "*X-tte*", X is extracted as a focus expression. The system multiplies the score, ($Score(d)$), of the candidate paragraph having "*X-wa*", "*X-towa* or something by 1.1. When the candidate expression includes focus expressions having modifiers (including modifier clauses and modifier phrases), the modifiers are used as candidate expressions, and the scores of the candidate expressions are multiplied by 1.1.

Below is an example of a candidate expression that is a modifier clause in a sentence.

Table 1: Results

Method	Correct	A	B	C	D
Method 1	57	18	42	10	89
Method 2	77	5	67	19	90

(There were a total of 100 questions.)

Question sentence: *sekai isan jouyaku to wa dono youna jouyaku desu ka?*

(What is the Convention concerning the Protection of the World Cultural and Natural Heritage?)

Sentence including answers:

<u>*1972 nen no dai 17 kai yunesuko soukai de saitaku sareta*</u> *sekai isan jouyaku ....*

(Convention concerning the Protection of the World Cultural and Natural Heritage, which <u>was adopted in 1972 in the 17th general assembly meeting of the UN Educational, Scientific and Cultural Organization.</u>)

Finally, our system extracts candidate expressions having high scores, ($Score(d)$s), as the preferred output. Our system extracts candidate expressions having scores that are no less than the highest score multiplied by 0.9 as the preferred output.

We constructed the methods for answer detection by consulting the dry run data in QAC-4.

4 Experiments

The experimental results are listed in Table 1. One hundred non-factoid questions were used in the experiment. The questions, which were generated by the QAC-4 organizers, were natural and not generated by using target documents. The QAC-4 organizers checked four or fewer outputs for each question. Methods 1 and 2 were used to determine what we used as answer candidate expressions (Method 1 uses one paragraph as a candidate answer. Method 2 uses one paragraph, two paragraphs, or three paragraphs as candidate answers.).

"A," "B," "C," and "D" are the evaluation criteria. "A" indicates output that describes the same content as that in the answer. Even if there is a supplementary expression in the output, which does not change

the content, the output is judged to be "A." "B" indicates output that contains some content similar to that in the answer but contains different overall content. "C" indicates output that contains part of the same content as that in the answer. "D" indicates output does not contain any of the same content as that in the answer. The numbers for "A," "B," "C," and "D" in Table 1 indicate the number of questions where an output belongs to "A," "B," "C," and "D". "Correct" indicates the number of questions where an output belongs to "A," "B," or "C". The evaluation criteria "Correct" was also used officially at NTCIR-6 QAC-4.

We found the following.

- Method 1 obtained higher scores in evaluation A than Method 2. This indicates that Method 1 can extract a completely relevant answer more accurately than Method 2.

- Method 2 obtained higher scores in evaluation "Correct" than Method 1. The rate of accuracy for Method 2 was 0.77 according to evaluation "Correct". This indicates that Method 2 can extract more partly relevant answers than Method 1. When we want to extract completely relevant answers, we should use Method 1. When we want to extract more answers, including partly relevant answers, we should use Method 2.

- Method 2 was the most accurate (0.77) of those used by all eight participating teams. We could detect paragraphs as answers including input terms and the key terms related to answer types based the methods discussed in Section 3.3. Our system obtained the best results because our method of detecting answers was the most effective.

Below is an example of the output of Method 1, which was judged to be "A."

Question sentence:

jusei ran shindan wa douiu baai ni okonawareru noka?

(When is amniocentesis performed on a pregnant woman?)

System output:

omoi idenbyou no kodono ga umareru no wo fusegu.

(To prevent the birth of children with serious genetic disorders)

Examples of answers given by organizers:

omoi idenbyou

(A serious genetic disorder)

omoi idenbyou no kodomo ga umareru kanousei ga takai baai

(To prevent the birth of children with serious genetic disorders.)

5 Conclusion

We constructed a system for answering non-factoid Japanese questions. An example of a non-factoid question is "Why are people opposed to the Private Information Protection Law?" We used various methods of passage retrieval for the system. We extracted paragraphs based on terms from an input question and output them as the preferred answers. We classified the non-factoid questions into six categories. We used a particular method for each category. For example, we increased the scores of paragraphs including the word "reason" for questions including the word "why." We participated at NTCIR-6 QAC-4, where our system obtained the most correct answers out of all the eight participating teams. The rate of accuracy was 0.77, which indicates that our methods were effective.

We would like to apply our method and system to Web data in the future. We would like to construct a sophisticated system that can answer many kinds of complicated queries such as non-factoid questions based on a large amount of Web data.

Acknowledgements

References

Yoshiaki Asada. 2006. Processing of definition type questions in a question answering system. *Master's thesis, Yokohama National University.* (in Japanese).

Adam Berger, Rich Caruana, David Cohn, Dayne Freitag, and Vibhu Mittal. 2000. Bridging the lexical chasm: Statistical approaches to answer-finding. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR-2000)*, pages 192–199.

Sasha Blair-Goldensohn, Kathleen R. McKeown, and Andrew Hazen Schlaikjer. 2003. A hybrid approach for qa track definitional questions. In *Proceedings of the 12th Text Retrieval Conference (TREC-2003)*, pages 185–192.

Charles L. A. Clarke, Gordon V. Cormack, and Thomas R. Lynam. 2001. Exploiting redundancy in question answering. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Susan Dumis, Michele Banko, Eric Brill, Jimmy Lin, and Andrew Ng. 2002. Web question answering: Is more always better? In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Kyoung-Soo Han, Young-In Song, Sang-Bum Kim, and Hae-Chang Rim. 2005. Phrase-based definitional question answering using definition terminology. In *Lecture Notes in Computer Science 3689*, pages 246–259.

Abraham Ittycheriah, Martin Franz, Wei-Jing Zhu, and Adwait Ratnaparkhi. 2001. IBM's Statistical Question Answering System. In *TREC-9 Proceedings*.

Julian Kupiec. 1993. MURAX: A robust linguistic approach for question answering using an on-line encyclopedia. In *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Hideyuki Maehara, Jun'ichi Fukumoto, and Noriko Kando. 2006. A BE-based automated evaluation for question-answering system. *IEICE-WGNLC2005-109*, pages 19–24. (in Japanese).

Bernardo Magnini, Matto Negri, Roberto Prevete, and Hristo Tanev. 2002. Is it the right answer? Exploiting web redundancy for answer validation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*.

Yuji Matsumoto, Akira Kitauchi, Tatsuo Yamashita, Yoshitaka Hirano, Hiroshi Matsuda, and Masayuki Asahara. 1999. Japanese morphological analysis system ChaSen version 2.0 manual 2nd edition.

Dan Moldovan, Marius Pasca, Sanda Harabagiu, and Mihai Surdeanu. 2003. Performance issues and error analysis in an open-domain question answering system. *ACM Transactions on Information Systems*, 21(2):133–154.

Kokoro Morooka and Jun'ichi Fukumoto. 2006. Answer extraction method for why-type question answering system. *IEICE-WGNLC2005-107*, pages 7–12. (in Japanese).

Masaki Murata, Kiyotaka Uchimoto, Hiromi Ozaku, Qing Ma, Masao Utiyama, and Hitoshi Isahara. 2000. Japanese probabilistic information retrieval using location and category information. *The Fifth International Workshop on Information Retrieval with Asian Languages*, pages 81–88.

Masaki Murata, Masao Utiyama, Qing Ma, Hiromi Ozaku, and Hitoshi Isahara. 2001. CRL at NTCIR2. *Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization*, pages 5–21–5–31.

Masaki Murata, Qing Ma, and Hitoshi Isahara. 2002. High performance information retrieval using many characteristics and many techniques. *Proceedings of the Third NTCIR Workshop (CLIR)*.

National Institute of Informatics. 2002. *Proceedings of the Third NTCIR Workshop (QAC)*.

S. E. Robertson and S. Walker. 1994. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. 1994. Okapi at TREC-3. In *TREC-3*.

Radu Soricut and Eric Brill. 2004. Automatic question answering: Beyond the factoid. In *In Proceedings of the Human Language Technology and Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL-2004)*, pages 57–64.

TREC-10 committee. 2001. The tenth text retrieval conference. http://trec.nist.gov/pubs/trec10/t10_proceedings.html.

Jinxi Xu, Ana Licuanan, and Ralph Weischedel. 2003. TREC 2003 QA at BBN: answering definitional questions. In *Proceedings of the 12th Text Retrieval Conference (TREC-2003)*, pages 98–106.