Answering any class of Japanese non-factoid question by using the Web and example Q&A pairs from a social Q&A website

Tatsunori Mori Mitsuru Sato* Madoka Ishioroshi
Graduate School of Environment and Information Sciences
Tokiwadai 79-7, Hodogaya-ku, Yokohama 240-8501, Japan
{mori,mitsuru,madoka}@forest.eis.ynu.ac.jp

Abstract

In this paper, we propose a method of non-factoid Web question-answering that can uniformly deal with any class of Japanese non-factoid question by using a large number of example Q&A pairs. Instead of preparing classes of questions beforehand, the method retrieves already asked question examples similar to a submitted question from a set of Q&A pairs. Then, instead of preparing clue expressions for the writing style of answers according to each question class beforehand, it dynamically extracts clue expressions from the answer examples corresponding to the retrieved question examples. This clue expression information is combined with topical content information from the question to extract appropriate answer candidates.

The experimental results showed that the clue expressions obtained from the set of examples improved the accuracy of answer candidate extraction.

1. Introduction

Question-answering (QA) technology is widely regarded as an advancement that combines information retrieval (IR) and information extraction (IE). QA systems do not provide us with relevant documents; instead, they directly provide answers to questions.

With regard to information sources, many recent studies have focused on Web documents because the Web is an up-to-date information resource. From the viewpoint of the style of answers, question answering tasks are classified into two categories: factoid type QA, in which answers are supposed to be short expressions like names and numerical expressions, and non-factoid type QA, in which answers are supposed to be relatively long descriptive expressions like definitions and causes. In this paper, we deal with Japanese non-factoid QA for Web documents.

Non-factoid questions can be divided into several classes in terms of the content of their answers, for example, definition-type, why-type, how-type, and so on. Since the clue expressions for answers are usually peculiar to each question class, many previous studies employed question classifiers to determine classes, and then applied one of several answer extraction methods, each of which was specific to a question class. While classes of factoid questions can be defined according to the categories in a thesaurus, classes of non-factoid questions are not well defined. With the exceptions of some typical classes like the definition-type, why-type, and how-type, it is difficult to distinguish and define all classes comprehensively.

In order to deal with this issue, in this paper, we propose a method to utilize a large number of example question-and-answer (Q&A) pairs from a social Q&A website. Instead of preparing classes of questions beforehand, this method retrieves already asked question examples that are similar to a submitted question from the set of Q&A pairs. Then, based on the writing style of the answers, it dynamically extracts clue expressions from the answer examples that correspond to the retrieved question examples. This clue expression information is combined with topical content information from the question to extract appropriate answer candidates. Note that we utilize the set of Q&A pairs, not to find answers from them, but to obtain clue expressions about the writing style of their answers. The information source for question answering is the Web documents retrieved by using the API of a Web search engine.

2. Related studies

2.1. Outline for answering non-factoid questions

Table 1 shows typical classes of non-factoid questions and Japanese examples of the writing styles of questions and answers. Some fixed expressions are observed in both questions and answers according to the class of the question.

Answer candidates for such non-factoid questions tend to be descriptive expressions, which are relatively long and cover a series of sentences. As described by Han et al.[3] with regard to definitional question-answering, the appropriateness of such relatively long answer candidates can be estimated by the combination of, at least, the following two measures.

Measure 1: Relevance to the topic of the question, how relevant is the candidate to the topic of the question?

[0]His current affiliation is Yahoo Japan Corporation.

Table 1. Typical classes of non-factoid questions

Class of questions	Examples of typical writing style	
	Question	Answer
Definition-type	∼*-tte-nani* (What is ∼)	∼*-towa* · · ·*-dearu* (∼ is · · ·)
Why-type	*Naze* ∼ (Why ∼)	· · · *tame* (Because · · ·)
How-type	∼*-suru-niwa dou-shitara ii* (How can I do ∼)	∼*-suru-niwa mazu* · · · (In order to do ∼, · · ·)
Other types	X-*to* Y-*no chigai-wa nani* (What is the difference between X and Y)	X-*wa* ∼*-daga*, Y-*wa* · · · (While X is ∼, Y is · · ·)

Measure 2: Appropriateness of writing style, how well does the candidate satisfy the writing style that is appropriate for answers of the class of the given question?

Here, by the term "writing style," we refer to the style of expressions peculiar to a class of questions and their answers, as shown in Table 1. Measure 1 can be implemented as the content similarity between a given question and an answer candidate. In many previous studies, Measure 2 was estimated according to the application results of rules that detected certain writing styles. Such a set of rules was realized as a set of manually constructed lexico-syntactic patterns, or a classifier that was generated from a set of example answers[6].

2.2. Related studies that treated questions on a class-by-class basis

There are many studies that classified non-factoid questions into classes, then treated questions on a class-by-class basis [7, 1, 5]. Measure 2 was usually implemented class-by-class. There have also been many studies that dealt with one specific question class.

Han et al.[3] proposed a probabilistic model for definitional question-answering. To determine Measure 1, the probabilistic model was constructed from top-ranked documents retrieved by query using the question target. The model for Measure 2 was obtained from a corpus of definitions.

2.3. Related studies that did not need question classification

Since, as described before, the classes of non-factoid questions are not well defined, it is difficult to distinguish and define all classes comprehensively. Moreover, the accuracy of a question classifier affects the overall accuracy of question-answering, because misclassified questions are incorrectly routed to an answering module for a different class. Therefore, a method is needed that does not depend on question classification and can deal with any class of questions uniformly.

Mizuno et al[4] proposed a method to realize Measure 2 without the classification of questions in Japanese non-factoid question answering. Using example Q&A pairs from a social Q&A website, it learns a binary classifier that judges whether or not the class of a given answer candidate is consistent with the class of a given question. By using this classifier, Measure 2 is realized without question classification. Because of the nature of the method, the length of the answer candidates should be predetermined as some text unit, like a paragraph. Therefore, the length of answer candidates is fixed and cannot be changed dynamically for a given question. With regard to the preparation of training data, negative examples should be artificially created by combining questions with answers from other questions.

Soricut et al.[9] also proposed an English non-factoid question answering system that does not need question classification. They introduced a statistical translation model between questions and the corresponding answers in order to bridge the lexical gap between questions and answers. A set of example Q&A pairs from FAQ sites on the Web are used for the estimation of the model. The model makes no distinction between the probability in terms of Measure 1 and that of Measure 2. Therefore, a large number of FAQs is supposed to be needed in order to guarantee the coverage of content words in questions. In this model, the length of the answers should be predetermined. Moreover, it requires a model that estimates the length of an answer from the length of the question.

2.4. Contribution of our proposed method

As answers for Japanese non-factoid questions, our proposed method extracts passages from Web documents that have higher values in terms of both Measure 1 and Measure 2. We also utilize example Q&A pairs from a social Q&A website in order to realize Measure 2, as did Mizuno et al.[4] and Soricut et al.[9].

However, our proposed method is different from these previous methods in terms of the usage of example Q&A pairs. Our method dynamically generates clue expressions from answer examples that correspond to question examples that are similar to the question submitted by a user. It is based not on a machine learning approach, but on an information retrieval approach. Therefore, it does not need a time-consuming learning process when new example Q&A pairs become available. There is no need to prepare unnatural negative examples as did Mizuno et al.[4]. The method used by Soricut et al.[9] obtained information about both Measure 1 and Measure 2 from Q&A examples, and therefore needed a wide-variety of examples in terms of genres and domains. On the other hand, since our method merely extracts clue expressions related to only Measure 2, we do not need to be concerned about the coverage of examples in

terms of topical content.

Our method also has the advantage of being able to adaptively determine the length of an answer candidate according to a submitted question.

3. Proposed method

Our proposed method does not classify the submitted questions at all. Instead of classification, the method directly retrieves example questions that are similar to a submitted question in terms of writing style, from a large collection of example Q&A pairs. Clue expressions in terms of the writing style of the answers are dynamically extracted from answer examples that correspond to retrieved question examples. These clue expressions are utilized for the estimation of Measure 2. This example-based method is expected to have the following advantages:

- It is free from the danger of question classification failure.
- Since extracted clue expressions are specific to not just a class of question but the submitted question itself, the clue expressions are more specialized and, therefore, expected to contribute to finding answer candidates that are more suitable to the question.
- Since the Q&A pairs in this example are written in a colloquial style, clue expressions extracted from them are expected to be suitable for extracting answer candidates from Web pages, many of which are also written in a colloquial style.

Figure 1 shows an outline of our proposed method.

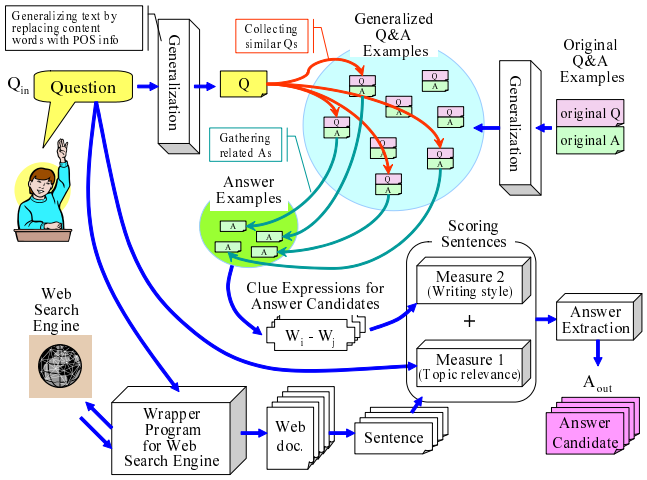

Figure 1. Outline of the proposed method

3.1. Obtaining clue expressions from Q&A examples

3.1.1. Q&A examples

We utilized a corpus of Q&A examples submitted to "Yahoo! Chiebukuro," which is a social Q&A website and the Japanese version of "Yahoo! answers." This corpus included about 3.1 million questions and 13.5 million answers that were contributed during the period from April 2004 to October 2005. Although each question had multiple answers, we utilized only the "best answers," which were selected as the best by the questioners. Hereafter, we use the term "Q&A pair" to refer to a pair consisting of a question and its best answer.

Since answers that include URLs tend to offer reference information only, we excluded such answers. In general, an answer text consisted of multiple sentences, some of which may have expressed information irrelevant to the answer. Since the essentials of an answer text tended to appear in the first half, we only adopted the first half of each answer text when the answer text contained more than one sentence. There are many questions that do not contain interrogatives, but we substituted the expression "∼*-wa nan-desu-ka* (what is ∼)" for some frequently appearing expressions like "∼*-o oshie-te* (please teach me ∼)," "∼*-o shiri-tai* (I want to know ∼)," and so on.

3.1.2. Generalizing texts in Q&A pairs

In order to extract only information about the writing style from examples for Measure 2, in this stage, we applied the following generalization to question texts and answer texts in the set of Q&A pairs. After word segmentation[1], the functional words, like interrogatives, postpositional particles, and so on, and a set of special content words described later are left as they are[2]. On the other hand, other words are replaced with their part-of-speech names. The set of special content words includes a set of Japanese content words that tend to be the foci of questions, like "*riyuu* (reason)," "*houhou* (method)," "*imi* (meaning)," "*chigai* (difference)," and so on. Verbs and adjuncts that appear with high frequency are also included in the set of special content words.

3.1.3. Generalizing question texts

From an examination of 30 Japanese sample questions from the evaluation workshop NTCIR-6 QAC[2], we found that a word 7-gram whose center word is the interrogative of the question generally provides enough information to determine the class of question. Therefore, after the generalization described in Section 3.1.2, at this stage each example question was replaced with a 7-gram extracted from the question as a further generalization.

After deleting Q&A pairs on the condition that the question text did not have the interrogative or the 7-gram extracted from the question text had a low frequency [3] in the set of all Q&A pairs, we obtained 0.9 million Q&A pairs.

For example, from the following sample question (1-J), whose English translation is (1-E):

(1) J. *shouhi zeikomi- no nedan- ga hyouji- sa- reru you- ni nat- ta riyuu- wa nan- desu- ka.*

E. Why do prices including the consumption tax come to be used for indicating prices?

we obtained the following 7-gram as a generalization:

[1]Since Japanese sentences do not have word boundaries explicitly, we need to apply word segmentation to sentences by using a morphological analyzer.

[2]More precisely, they are replaced with the strings that represent their pronunciations. Some Japanese characters are not phonograms but ideograms. Therefore, replacing a word with its pronunciation is one of generalizations that cope with the diversity of variants

[3]In this paper, we remove 7-grams that appear less than three times.

(2) *ta riyuu wa nan desu ka* ⟨period⟩

where "*nan* (what)" is the interrogative and the notation $\langle \cdots \rangle$ represents a part-of-speech name.

3.1.4. Retrieving example questions similar to the submitted question

In order to obtain clue expressions peculiar to answer candidates for the question submitted by a user, in this stage, the proposed method retrieves example Q&A pairs whose questions are similar to the submitted question from the viewpoint of writing style. As described before, a word 7-gram whose center word is an interrogative seems to give us enough context to determine the class of question. Therefore, we define the similarity between two questions as the similarity between the word 7-grams extracted from questions. In this paper, as the similarity between questions, we adopt a simple measure, i.e. the number of shared words in 7-grams. Since the interrogative plays a very important role in a question, we suppose that the similarity is zero when the interrogatives of two questions are different. According to the similarity, N-best example Q&A pairs are obtained by using an ordinary information retrieval technique.

For example, when the following Japanese question (3-J), whose English translation is (3-E), is submitted by a user,

(3) J. *Usu-zan no funka yochi ga seikou shi ta riyuu wa nan desu ka.*

E. What is the reason for success in the eruption predictions for Mt. Usu ?

we obtain the following 7-gram from the question:

(4) *ta riyuu wa nan desu ka* ⟨period⟩

By using the 7-gram as a query, some appropriate example Q&A pairs, including the pair for Question (1), are expected to be retrieved from the entire set of example Q&A pairs.

3.1.5. Extracting clue expressions from answer examples

In this stage, clue expressions are extracted from the answers in the example Q&A pairs obtained in the stage described in Section 3.1.4. In this paper, we adopted a 2-gram as a clue expression unit because it is the smallest unit that can represent relations between words. Here, we assume that the effectiveness of each 2-gram as a clue expression can be estimated by the degree of correlation between the 2-gram and the answers from the collected Q&A pairs.

As the measurement of the correlation, we adopted the χ^2 value shown in Equation (1) for the following two kinds of events for the answers from the entire set of example Q&A pairs:

event α Being an answer example that corresponds to one of the collected question examples, which are similar to the submitted question. The set of answer examples for the event is denoted by A.

event $\beta(b)$ Being an answer example that contains a certain 2-gram b. The set of answer examples for the event is denoted by $B(b)$.

$$\chi^2(b) = \frac{n}{|A| \cdot |\overline{A}| \cdot |B(b)| \cdot |\overline{B(b)}|} \cdot (|A \cap B(b)| \cdot |\overline{A} \cap \overline{B(b)}| - |\overline{A} \cap B| \cdot |A \cap \overline{B}|)^2 \quad (1)$$

where n is the total number of example Q&A pairs. The more correlated two events are, the larger the value of χ^2 is. According to the value of $\chi^2(b)$, the M-best 2-grams are selected as clue expressions of the answers for the submitted question.

For example, let us consider the situation where a user submits Question (3) to the system. Here, we assume that the question example (1) in Section 3.1.3 is paired with the following answer example:

(5) J. *omotemuki- wa nedan- o wakari yasuku suru tame.*

E. As for the ostensible reason, ϕ makes prices simple.

In this situation, some example questions similar to the submitted question, probably including Question (1), are collected in the stage described in Section 3.1.4. Simultaneously, answer examples corresponding to the example questions, including Answer (5), are also obtained. From the set of answer examples, we can obtain the M-best 2-grams with their χ^2 values as shown in Table 2 by using Equation (1).

Table 2. Examples of clue 2-grams

Clue 2-gram	χ^2 value
ta riyuu (reason for)	705
ta kara (because)	531
riyuu wa (the reason is)	219
⋮	⋮
kara ⟨period⟩ (because)	113
kara desu (because)	98
⋮	⋮

3.2. Question Answering using clue expressions obtained from Q&A examples

3.2.1. Extracting keywords from a question and obtaining their related words

From a question submitted by a user, content words are extracted as keywords. Let K, K_n, and K_p be the set of all keywords, the set of keywords of simple nouns (one-morpheme words), and the set of keywords except nouns, respectively. Since sequences of simple nouns may form compound nouns, let K_c be the set of all compound nouns and other remaining simple nouns.

A question usually contains only a few keywords and these may not be enough to estimate Measure 1. Therefore, the following keyword expansion and weighting are performed by using Web documents.

1. Create all subsets that contain three words from K_c.
2. Form a Boolean "AND" query q_i from each subset and submit it to a Web search engine to obtain a set of snippets. Let n_i be the number of obtained snippets.
3. The weight value $T(w_j)$ defined as the following equation is calculated for each word w_j in snippets:

$$T(w_j) = \max_i \frac{freq(w_j, i)}{n_i} \quad (2)$$

where $freq(w_j, i)$ is the frequency of the snippets that contain the word w_j for the query q_i.

In order to give each keyword $k \in K$ a weight value that is not less than those of the expanded words, the weight value is defined as the following equation:

$$T(k) = \max_j T(w_j) \quad (3)$$

For example, in the case of Question (3) in Section 3.1.4, we can obtain the related words shown in Table 3 from the Web.

Table 3. Obtained related words and their weight values

Related word	Weight value
kazan (volcano)	0.72
miyake-jima (Miyake Island)	0.29
renraku (coordinating)	0.29
kai (committee)	0.29
⋮	⋮

3.2.2. Retrieving Web documents

A Boolean "AND" query is formed using words in the set K. Then it is submitted to a Web search engine to obtain Web documents. The same procedure is also applied to the sets K_c and $K_c \cup K_p$, and further documents are obtained. Plain text is extracted from each Web document by some simple text processing, including the removal of HTML tags from the document.

3.2.3. Grading sentences in retrieved documents

In this stage, each sentence in the retrieved documents is graded in terms of both Measure 1 and Measure 2. First, by using the method in Section 3.1.4, the system collects example Q&A pairs whose questions are similar to the submitted question in terms of writing style. Second, by using the method described in Section 3.1.5, it extracts a set of 2-grams as clue expressions from the answer examples of the example Q&A pairs and calculates the corresponding $\chi^2(b)$ value for each 2-gram b. Finally, the score of each sentence is calculated by using the following equation:

$$\mathrm{Score}(S_i) = \frac{1}{\log(1+\mathrm{length}(S_i))} \cdot \left\{\sum_{j=1}^{n} T(w_{i,j})\right\}^{\gamma} \cdot \left\{\sum_{k=1}^{m} \sqrt{\chi^2(b_{i,k})}\right\}^{1-\gamma} \quad (4)$$

where n is the number of different words in the sentence S_i, m is the number of different 2-grams in S_i, $w_{i,j}$ is the j-th word in sentence S_i, and $b_{i,k}$ is the k-th 2-gram in S_i. Since the terms $\sum_{j=1}^{n} T(w_{i,j})$ and $\sum_{k=1}^{m} \sqrt{\chi^2(b_{i,k})}$ in Equation (4) correspond to Measure 1 and Measure 2, respectively, the parameter γ is used to determine the mixture ratio of Measure 1 and Measure 2. The normalization term $\frac{1}{\log(1+\mathrm{length}(S_i))}$ is introduced to calculate the density of content words related to the question (i.e. keywords and their related words) and clue expressions (i.e. 2-grams that correlate with answer examples). In order to reward longer sentences, a sentence length logarithm is adopted.

3.2.4. Extracting answer candidates

An outline of the extraction of answer candidates is shown in Figure 2. First, all sentences with maximal (not maximum) scores are selected from the Web documents retrieved in the stage described in Section 3.2.2. These play a role as the seeds of answer candidates. In this paper, an answer candidate corresponding to a seed is defined as the longest series of sentences that satisfies the following conditions: 1) the seed is in the series, and 2) every sentence in the series has a score greater than a threshold. The threshold is calculated seed by seed from a predetermined ratio in terms of score. For example, in Figure 2, the ratio is set as 50% of the maximal score. We define the score of an answer candidate as the maximal score.

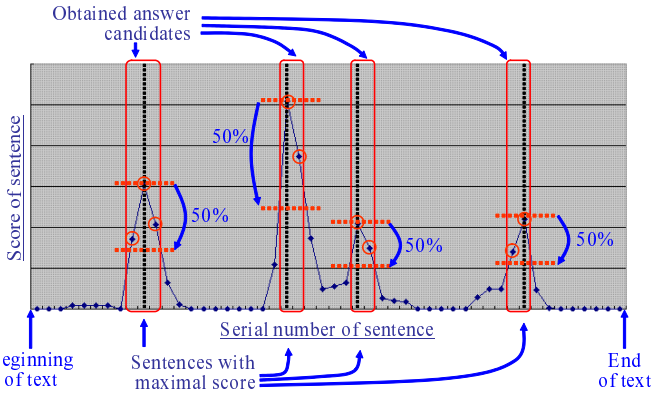

Figure 2. Extraction of answer candidates

As final answers, answer candidates are output in the descending order of their scores. For example, answer candidates, including the following example answer (6-J), whose English translation is (6-E), are obtained for Question (3).

(6) J. *Funka yochi no seikou-rei to shite-wa 2000-nen-no Usuzan-no funka-ga yuumei-de, 3-gatsu 27-nichi-kara-no kazan-sei jishin-no bunseki-ya dansou-no tansaku-ni yotte, kinjitsu-chu-no funka-ga-yochi-sare, futsuka-go-ni-wa kishouchou-kara kinkyu-kazan-jouhou-ga da-sare-mashita.*

E. The case of the eruption of Mt. Usu in 2000 is a famous successful example of eruption prediction. From the analysis of the volcanic earthquake and the investigation of the fault since March 27, an eruption within a couple of days was predicted. Two days later, the Meteorological Agency announced that it had urgent volcano information.

4. Experimental evaluation

We conducted an experiment for the evaluation of the effectiveness of our proposed method, by using the questions in the test set of the formal run of NTCIR-6 QAC[2], which is an evaluation workshop for non-factoid question answering. It contains 100 questions, and consists of 21 definitional questions, 33 why-type questions, 6 how-type questions, and 40 other-type questions. Note that a set of

newspaper articles are used as the information source in the evaluation of NTCIR-6 QAC. On the other hand, we used Web documents as the information source in this study. Therefore, we cannot directly compare the system based on our proposed method with the systems that participated in NTCIR-6 QAC.

4.1. Experiment

We compared the system based on the proposed method with two kinds of baseline systems.

Proposed method It takes account of both Measure 1 and Measure 2 evenly by using Equation (4) with the parameter setting $\gamma = 0.5$.

Baseline 1 It takes account of Measure 1 (topic relevance) only by using Equation (4) with the parameter setting $\gamma = 1.0$.

Baseline 2 It performs the detection of question classes. When the question class is the why-type or definition-type, it applies a set of lexico-syntactic patterns[4] to each sentence in order to detect writing styles peculiar to answers for the question class. Each pattern has a score with which the answer candidates are rewarded when the pattern is applicable. The score is added to the score of Baseline 1.

With regard to a Web search engine, we utilized the API for the search engine of Yahoo! JAPAN[5]. As described in Section 3.2.2, three kinds of queries were generated and submitted to the engine to obtain the top 50 documents for each query. However, the average number of different documents was only about 40 in the experiment, because some documents were redundant.

In the stage of keyword expansion described in Section 3.2.1, we obtained the top 100 snippets (i.e. $n_i = 100$). The number of retrieved Q&A examples described in Section 3.1.4 and the number of clue 2-grams described in Section 3.1.5 were 500 (i.e. $N = 500$) and 200 (i.e. $M = 200$), respectively.

In this experiment, the top five answers from each system were evaluated. The correctness was judged by the second author. When an answer candidate was a part of a correct answer or the candidate included a correct answer, the candidate was judged to be a *good* answer. With regard to evaluation measures, we adopted the mean reciprocal rank (MRR) and the number of well-answered questions, for which the system could find at least one good answer within the top 5 candidates. The reciprocal rank (RR) is the inverse of the rank of the first good answer for each question, and MRR is the average of the RRs for all the questions.

4.2. Experimental results

The experimental results are shown in Table 4. Although the proposed method and baseline 1 did not perform any question classification, the results are shown on a class-by-class basis in order to investigate the effectiveness of the method for each question class.

[4]Rules for question classification and lexico-syntactic patterns were manually prepared for another system that participated in NTCIR-6 QAC.

[5]http://www.yahoo.co.jp/

5. Discussion

First, the overall results in Table 4 show that the proposed method outperformed the two baselines. This means that clue 2-grams and their χ^2 values obtained from the set of example Q&A pairs effectively contributed to finding appropriate answer candidates. However, the accuracy was not largely improved for the definition-type questions. This was primarily because the words that appeared in the definition description tend to be easily collected by using the keyword expansion described in Section 3.2.1, without the need to use Measure 2. On the other hand, the clue 2-grams worked effectively for questions other than the definition-type questions.

By comparing Baseline 2 with Baseline 1, we can see that the accuracy was degraded for the questions of the how-type and other types, although Baseline 2 did not use any lexico-syntactic patterns for these question types. This was because there were some faults in the question classification. On the other hand, the lexico-syntactic patterns for detecting writing styles worked for questions of the definition-type and the why-type.

The following sentences show a successful example. When a user submits Question (7-J), whose English translation is (7-E), the system based on the proposed method outputs a good answer (8-J), whose English translation is (8-E).

(7) J. *sukeruton-to ryuuju-no chigai-wa nan-desu-ka.*

E. What is the difference between skeleton and luge?

(8) J. *ryuuju-wa ashi-o mae-ni suru-ga, sukeruton-wa atama-o mae-ni shi-ta harabai-no shisei-de koori-no koosu-o suberu.*

E. While in luge (racers lie) feet first, in skeleton (racers) slide on an ice-covered course in a head-first and face-down position.

6. Failure analysis

There were 35 questions that could not be answered correctly with the proposed method. Table 5 shows the failed processing stages and their number of questions. One of the advantages of our proposed method is that there were fewer errors in the earlier stages of question answering, because the method does not need question classification. There were three errors in the keyword extraction stage because of a failure in the segmentation of sentences by the morphological analyzer.

Table 5. Failure analysis

Stage failed	# of Q.
Keyword extraction from Q.	3/35
Document retrieval	9/35
Extraction of answer candidates	23/35

The number of failures in the document retrieval stage was not small. When documents that included correct answer candidates were not retrieved, the system could not find candidates even if the system achieved a very high performance in the answer extraction stage. The method of query formulation from questions should be refined.

Table 4. Experimental result

	Proposed ($\gamma = 0.5$)		Baseline 1 ($\gamma = 1.0$)		Baseline 2 (with patterns)	
Question type	MRR	# of well-answered Q.	MRR	# of well-answered Q.	MRR	# of well-answered Q.
Def-type	0.643	16/21	0.561	17/21	0.581	16/21
Why-type	0.307	16/33	0.178	11/33	0.194	11/33
How-type	0.297	4/6	0.067	2/6	0.056	1/6
Others	0.513	29/40	0.343	24/40	0.316	23/40
All	0.459	65/100	0.318	54/100	0.316	51/100

An example case of failure in the extraction of answer candidates for Question (9-J) is shown in (10-J). Their English translations are shown in (9-E) and (10-E), respectively.

(9) J. *rinkai-to-wa dono-youna joutai-no koto de-su-ka.*

E. What state is criticality?

(10) J. *Chou-rinkai-to-wa ekitai-joutai-de-mo kitai-joutai-demo-nai joutai-de, atsuryoku, ondo-ga rinkai-ten-o koeta joutai-o ippanteki-ni chou-rinkai-ryuutai-to yobi-masu.*

E. Super-criticality is the state of being neigher a fluid nor a gas, and the state at a pressure and temperature above the critical point is generally called a supercritical fluid.

As shown in this example, there are cases in which answer candidates are not suitable for the answer for a submitted question, even if the answer candidate includes keywords, their related words, and clue 2-grams. In the case of (10), the constituent noun ("*rinkai* (criticality)") of the compound noun ("*chou-rinkai* (super-criticality)") is wrongly matched with the keyword ("*rinkai*") in the question. In order to cope with problems, including this type of case, we have to introduce a more sophisticated language model for answer descriptions.

7. Concluding remarks

In this paper, we proposed a method of Web question-answering that can deal with any class of non-factoid question uniformly by using a large number of example Q&A pairs. Instead of preparing classes of questions beforehand, the method retrieves already asked question examples similar to a submitted question from a set of Q&A pairs. Then, instead of preparing clue expressions for answers according to each question class beforehand, it dynamically extracts clue expressions from the answer examples corresponding to the retrieved question examples. This clue expression information is combined with topical content information from the question to extract appropriate answer candidates.

The experimental results showed that the clue expressions obtained from the set of examples improved the accuracy of the extraction of answer candidates. We also confirmed that there were some cases where the proposed method could not extract appropriate answer candidates. In order to deal with this issue, we have to introduce a more sophisticated language model for answer descriptions and a scoring function for answer candidates based on this model, like a model-based approach[8]. We would like to investigate such models in our future work.

Acknowledgment

We would like to thank Yahoo Japan Corporation and The National Institute of Informatics who provide us the data set of Yahoo! Chiebukuro. We are also grateful to the organizers of NTCIR-6 QAC and people who manage the NTCIR workshops. We would like to especially thank Mainichi Shimbun for permitting us to use the documents for research. Finally, we would like to thank the anonymous reviewers for their helpful comments.

This study was partially supported by Grant-in-Aid for Scientific Research on Priority Areas (No.19024033) from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

References

[1] J. Fukumoto. Question Answering System for Non-factoid Type Questions and Automatic Evaluation based on BE Method. In *Proceedings of the Sixth NTCIR Workshop*, pages 441–447, May 2007.

[2] J. Fukumoto, T. Kato, F. Masui, and T. Mori. An Overview of the 4th Question Answering Challenge (QAC-4) at NTCIR Workshop 6. In *Proceedings of the Sixth NTCIR Workshop Meeting*, pages 433–440, 5 2007.

[3] K.-S. Han, Y.-I. Song, and H.-C. Rim. Probabilistic model for definitional question answering. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 212–219, 2006.

[4] J. Mizuno, T. Akiba, A. Fujii, and K. Itou. Non-factoid Question Answering Experiments at NTCIR-6: Towards Answer Type Detection for Realworld Questions. In *Proceedings of the Sixth NTCIR Workshop*, pages 487–492, May 2007.

[5] M. Murata, S. Tsukawaki, T. Kanamaru, Q. Ma, and H. Isahara. Non-Factoid Japanese Question Answering through Passage Retrieval that Is Weighted Based on Types of Answers. In *Proceedings of the third IJCNLP*, Jan. 2008.

[6] Ryuichiro Higashinaka and Hideki Isozaki. Corpus-based Question Answering for why-Questions. In *Proceedings of the third IJCNLP*, Jan. 2008.

[7] H. Shima and T. Mitamura. JAVELIN III: Answering Non-Factoid Questions in Japanese. In *Proceedings of the Sixth NTCIR Workshop*, pages 464–468, May 2007.

[8] S. Sinha and S. Narayana. Model-based answer selection. In *Proceedings of AAAI-05 Workshop on Inference for Textual Quesiton Answering*, Jan. 2005.

[9] R. Soricut and E. Brill. Automatic Question Answering Using the Web: Beyond the Factoid. *Journal of Information Retrieval - Special Issue on Web Information Retrieval*, 9(2):191–206, Mar. 2006.