SpringerPlus
a SpringerOpen Journal

RESEARCH Open Access

Question answering system using Q & A site corpus Query expansion and answer candidate evaluation

Kanako Komiya[1*], Yuji Abe[1], Hajime Morita[2] and Yoshiyuki Kotani[1]

Abstract

Question Answering (QA) is a task of answering natural language questions with adequate sentences. This paper proposes two methods to improve the performance of the QA system using a Q&A site corpus. The first method is for the relevant document retrieval module. We proposed modification of measure of mutual information for the query expansion; we calculate it between two words in each question and a word in its answer in the Q&A site corpus not to choose the words that are not suitable.

The second method is for the candidate answer evaluation module. We proposed to evaluate candidate answers using the two measures together, i.e., the Web relevance score and the translation probability. The experiments were carried out using a Japanese Q&A site corpus. They revealed that the first proposed method was significantly better than the original method when their accuracies and MRR (Mean Reciprocal Rank) were compared and the second method was significantly better than the original methods when their MRR were compared.

Introduction

Question Answering (QA) is a task of answering questions written in natural language with adequate sentences, which consists of the following four modules (Soricut and Brill 2006).

(1) Question analysis
(2) Relevant document retrieval
(3) Candidate answer extraction
(4) Candidate answer evaluation

When a question written in natural language is input into the system, the system carries out keyword extraction in the question analysis module. Then the system retrieves relevant documents using the keywords that were obtained in the last module in the relevant document retrieval module. After that, the system extracts candidate answers in the candidate answer extraction module. The size of the candidate answers varies according to their question types, e.g., a phrase or a sentence. A sentence or a paragraph will be the candidate answer when the QA is non-factoid. Finally, the system estimates the qualities of the candidate answers that were obtained in the candidate answer extraction module in the candidate answer evaluation module.

This paper proposes two methods to improve the performance of the QA system using a Q&A site corpus. The first method is for the relevant document retrieval module. We proposed modification of measure of mutual information for the query expansion. The query expansion is an approach to extend query words by adding new words that are not included in each question to improve the qualities of the relevant documents to be retrieved. In previous work, words to be added are chosen based on mutual information between a word of each question and a word of its answer in the Q&A site corpus (Berger et al. 2000). We calculated it between two words in each question and a word in its answer in the Q&A site corpus not to choose the words that are not suitable for the query expansion.

The second method is for the candidate answer evaluation module. The QA system estimates the qualities of candidate answers that were obtained by the document retrieval in this module. This module is important because it directly affects system's outputs. There are

*Correspondence: kkomiya@cc.tuat.ac.jp
[1] Institute of Engineering, Tokyo University of Agriculture and Technology, 2-24-16 Nakacho-Kotanei, Tokyo, 184-8588, Japan
Full list of author information is available at the end of the article

two cues to estimate candidate answers, i.e., 1) the topic relevance, which evaluates association between each candidate answer and its question in terms of its content, and 2) the writing style, which evaluates how the writing style of each candidate answer corresponds to its question type. In this paper, we propose to evaluate candidate answers using the Web relevance score (Ishioroshi et al. 2009) and the translation probability (Soricut and Brill 2006) together.

We will show that our proposed methods improved each module by the experiments using a Japanese Q&A site corpus.

This paper is organized as follows. Section 'Related work' reviews related work on QA. Sections 'Query expansion using mutual information' and 'Query expansion using two words in a question' explain how words for query expansion were determined in the relevant document retrieval module in the previous work (Berger et al. 2000) and the first proposed method for the module, respectively. Sections 'Candidate answer evaluation' and 'Candidate answer evaluation with web relevance score and translation probability' describe how candidate answers were evaluated in the candidate answer module in the previous work (Ishioroshi et al. 2009 and Soricut and Brill 2006) and the second proposed method for the module. Section 'Experiments' explains the experimental settings. We present the results in Section 'Results' and discuss them in Section 'Discussion'. Finally, we conclude the paper in Section 'Conclusion'.

Related work

Question Answering (QA), which involves answering questions written in natural language with adequate sentences, has been studied intensively in recent years within or outside the area of natural language processing. The QA systems within the area are sometimes called as open domain question answering systems because they are not domain specific (Ishioroshi et al. 2009).

Types of questions that are treated by the QA systems can be categorized into two kinds, i.e., factoid and non-factoid. Questions of the former type ask the names of people or places, or the amounts of stuffs, e.g., "How tall is Mt. Fuji?". On the other hand, questions of the latter type ask definitions, reasons, or methods, e.g., "What are ES cells?". Our system treats the both types of questions in this paper.

We proposed two methods to improve the performance of the QA system; the first method is for the query expansion of the relevant document retrieval module and the second method is for the candidate answer evaluation module. For the query expansion, Saggion and Gaizauskas (2004) proposed to obtain words for the query expansion using relevance feedback from the Web. They regarded words that appeared frequently in documents retrieved for each question query as the new words for the query expansion. Mori et al. (2007) and Derczynski et al. (2008) used tf-idf and Lin et al. (2010) used Okapi-BM25 for the criteria instead of the term frequency of Saggion and Gaizauskas (2004). Lao et al. (2008) proposed to obtain the synonyms of words in each question using bootstrap method and to use them for the query expansion. Saggion and Gaizauskas (2004) also used synonyms but obtained them from a dictionary. Liu et al. (2010) obtained them from Wikipedia. Finally, Berger et al. (2000) proposed to learn what kind of words tend to appear in answers when some words appeared in questions using a Q&A site corpus and to use words that frequently appear for the query expansion. We improved one of the approaches suggested by Berger et al. (2000) in this paper.

For the query expansion, some researchers such as Higashinaka and Isozaki (2007,2008) and Isozaki and Higashinaka (2008) reported that the performance of the system improved when the question types were classified into classes such as "how-questions" and "why-questions" in advance. However, Ishioroshi et al. (2009) and Soricut and Brill (2006) developed a QA system without classification of the question types. Ishioroshi et al. (2009) estimated the topic relevance by relevance feedback from the Web.

Soricut and Brill (2006) and Berger et al. (2000) treated QA task as translation and succeeded in evaluating the topic relevance and the writing style simultaneously. We also improved them by combining their methods together without classification of the question types.

Query expansion using mutual information

Berger et al. (2000) proposed to learn what kind of words tend to appear in answers when some words appeared in questions using a Q&A site corpus and to use words that frequently appear for the query expansion. In their work, mutual information was used to measure the degree of relevance between a word in each question and a word in its answer. The formula of mutual information is as follows:

$$\begin{aligned} I(W_q, W_a) = {} & P(w_q, w_a) log \frac{P(w_q, w_a)}{P(w_q)P(w_a)} \\ & + P(w_q, \bar{w_a}) log \frac{P(w_q, \bar{w_a})}{P(w_q)P(\bar{w_a})} \\ & + P(\bar{w_q}, w_a) log \frac{P(\bar{w_q}, w_a)}{P(\bar{w_q})P(w_a)} \\ & + P(\bar{w_q}, \bar{w_a}) log \frac{P(\bar{w_q}, \bar{w_a})}{P(\bar{w_q})P(\bar{w_a})}, \end{aligned} \tag{1}$$

where W_q and W_a represent binary random variables that show whether a word w_q appear in each question and whether a word w_a appear in its answer, respectively.

$$W_q = \begin{cases} w_q & (w_q \textit{ appears in a question}) \\ \bar{w_q} & (w_q \textit{ dose not appear in a question}) \end{cases} \quad (2)$$

$$W_a = \begin{cases} w_a & (w_a \textit{ appears in an answer}) \\ \bar{w_a} & (w_a \textit{ dose not appear in an answer}) \end{cases} \quad (3)$$

The more w_q and w_a co-occur in a corpus, the grater their mutual information becomes.

Berger et al. (2000) chose a word from its answer for every words in each question. It was the word that maximized mutual information between the question word and the answer word itself. After this, {a word in a question → a word in an answer} denotes the query expansion using this method.

This method works effectively when the training and test corpora are domain specific[a]. However, it sometimes causes semantic drift when corpora are large and not domain specific. For example, when the question was "What are the connections between softbank and yahoo?", it gave us the following results: {softbank → hawks}[b] and {yahoo → mail}. *Hawks* and *mail* are relevant with *softbank* and *yahoo*, respectively, but they should not be used for the query expansion because they are no relevance with the original question.

Query expansion using two words in a question

In order to alleviate the semantic drift, we propose to use mutual information based on two words in each question and a word in its answer. The new equation of mutual information is as follows:

$$\begin{aligned} I(W_{q1}, W_{q2}, W_a) = {} & P(w_{q1}, w_{q2}, w_a) log \frac{P(w_{q1}, w_{q2}, w_a)}{P(w_{q1}, w_{q2})P(w_a)} \\ & + P(w_{q1}, w_{q2}, \bar{w_a}) log \frac{P(w_{q1}, w_{q2}, \bar{w_a})}{P(w_{q1}, w_{q2})P(\bar{w_a})} \\ & + P(\bar{w_{q1}}, \bar{w_{q2}}, w_a) log \frac{P(\bar{w_{q1}}, \bar{w_{q2}}, w_a)}{P(\bar{w_{q1}}, \bar{w_{q2}})P(w_a)} \\ & + P(\bar{w_{q1}}, \bar{w_{q2}}, \bar{w_a}) log \frac{P(\bar{w_{q1}}, \bar{w_{q2}}, \bar{w_a})}{P(\bar{w_{q1}}, \bar{w_{q2}})P(\bar{w_a})} \end{aligned} \quad (4)$$

It represents the degree of co-occurrence between two words in a question and a word in its answer. The more w_{q1}, w_{q2}, and w_a co-occur in a corpus, the grater their mutual information becomes like equ. (1). For example, when the question was "What are the connections between softbank and yahoo?" , it gave us the following results: {softbank and yahoo → subsidiary}.

Candidate answer evaluation

The second proposed method is for the candidate answer evaluation. As mentioned above, the topic relevance and the writing style are used to estimate candidate answers. We introduced two existing methods for the module. First method is the work proposed by Ishioroshi et al. (2009), which estimated the topic relevance by relevance feedback from the Web. They regarded words that frequently appeared in documents retrieved for each question query as relevant words. Therefore, candidate answers that contain many relevant words were regarded better in terms of the topic relevance.

The relevance words are obtained as follows:

(1) Make a keyword class K that contains content words (i.e., nouns, verbs, and adverbs) in each question
(2) Choose three words from K in all combinations and search the Web by them
(3) Obtain at most 100 Web snippets, i.e., summaries of the Web documents that were obtained by a Web search engine, for each query

Each content word w_j in these snippets is treated as a relevant word for the question. The relevance degree of the relevant word, i.e., $T(w_j)$ is defined by the following equation:

$$T(w_j) = \max_i \frac{freq(w_j, i)}{n_i}, \quad (5)$$

where i represents a index of a query (i.e., triple of content words), n denotes the number of snippets obtained from i_{th} query, $freq(w_j, i)$ denotes the number of snippets that contain word w_j that were obtained from i_{th} query. Candidate answer evaluation score in terms of the topic relevance, i.e., $Web\_relevance(Q, A)$ is defined as the sum of the relevant degrees of the relevant words contained in each candidate answer as follows:

$$Web\_relevance(Q, A) = \sum_{j=1}^{l} T(w_j), \quad (6)$$

where Q represents a question, A represents its candidate answer, l denotes the number of words in the candidate answer, and w_j denotes each word in the candidate answer.

Finally, Ishioroshi et al. (2009) evaluated candidate answers using the following score that took into consideration the topic relevance and the writing style as well:

$$\begin{aligned} Score(S_i) = {} & \frac{1}{ln(1 + length(S_i))} \\ & \times \left\{ \sum_{j=1}^{l} T(w_{i,j}) \right\}^{\gamma} \cdot \left\{ \sum_{k=1}^{m} \sqrt{\chi^2(b_{i,k})} \right\}^{1-\gamma}, \end{aligned} \quad (7)$$

where l denotes the number of types of word $w_{i,j}$ in a sentence S_i, m represents the number of types of writing feature $b_{i,k}$ in S_i, $length(S_i)$ means the number of the characters of S_i, χ^2 denotes the score of each writing style, and γ represents the weighting parameter. As for χ^2, a chi-square value is calculated between the answers that

include the writing feature $b_{i,k}$ and the top N answers retrieved for the question query.

The second method is the work by Soricut and Brill (2006), which treated QA task as translation. They succeeded in evaluating the topic relevance and the writing style simultaneously. In their method, each question and its answer are regarded as the source and target sentences, respectively. For translation, word-by-word translation probabilities are learned using a Q&A site corpus. When a question is input into the system, this system calculates the translation probabilities from the question into their candidate answers. Then the candidate answers are evaluated using their probabilities. They used the IBM-Model1 (Brown et al. 1993) as a translation model, which is simple but showed efficacy in many tasks. The answer evaluation is formulated as follows using IBM-Model1 (as in Berger et al. (2000)) :

$$A^{\star} = \underset{A}{\operatorname{argmax}}\, P(A|Q) = \underset{A}{\operatorname{argmax}}\, P(Q|A)P(A) \tag{8}$$

$$P(Q|A) = \epsilon \prod_{j=1}^{m} \left(\frac{l}{l+1}\left(\sum_{i=1}^{l} p(q_j|a_i) \cdot c(a_i|a)\right) + \frac{1}{l+1} p(q_j|NULL)\right), \tag{9}$$

where $A^{\star}$ represents the most adequate candidate answer, $Q(= q_1, q_2, \ldots, q_m)$ and $A(= a_1, a_2, \ldots, a_l)$ each represents a question and its candidate answer, m and l each denotes the number of words in the question and its candidate answer[c], $P(q|a)$ represents the translation probability from a word a in an answer to a word q in a question, $c(a_i|a)$ are the relative counts of the answer words, $P(A)$ denotes generation probability of the candidate answer A, and ϵ is a probability of generating a question whose length is m from the candidate answer.

We can have the equation (10) by assuming that $c(a_i|a)$ is $1/l$ like Brown et al. (1993).

$$P(Q|A) \approx \frac{\epsilon}{(l+1)^m} \prod_{j=1}^{m} \sum_{i=1}^{l+1} P(q_j|a_i) \tag{10}$$

In equation (10), there is a problem where the less the number of words in a candidate answer becomes, the more its translation probability increases because the value of the coefficient increases as l decreases. Therefore, we neglected the coefficient and got equation (11) instead of equation (10).

$$P(Q|A) \approx \prod_{j=1}^{m} \sum_{i=1}^{l+1} P(q_j|a_i) \tag{11}$$

Candidate answer evaluation with web relevance score and translation probability

When evaluating the topic relevance, the method using the translation probability proposed by Soricut and Brill (2006) can flexibly capture synonyms. This is because the translation probabilities are learned from the massive examples of a Q&A site corpus beforehand. However, it is unable to capture the co-occurrence information of several words in a question because it only utilizes word-to-word translation probabilities. By contrast, the Web relevance score proposed by Ishioroshi et al. (2009) can capture the co-occurrence information but cannot capture the synonyms because the Web documents dynamically obtained are small. Thus, it seems that the answer evaluation method using these methods simultaneously would be able to achieve the greater performance.

New answer evaluation formula

Equation (12) is the new formula of the answer evaluation that uses the Web relevance score and the translation probability:

$$EvalScore(Q,A) = \mathscr{P}(Q,A)^{1-\gamma} \cdot Web\_relevance(Q,A)^{\gamma} \tag{12}$$

$$\mathscr{P}(Q,A) = P(Q|A)P(A), \tag{13}$$

where $\mathscr{P}(Q,A)$ represents the probability that should be maximized in the equation (8) (the score by Soricut and Brill (2006)), $Web\_relevance(Q,A)$ denotes the score using Web relevance score (the score by Ishioroshi et al. (2009)), and γ represents the weighting parameter. The equation (12) is equivalent to the translation probability when $\gamma = 0$ whereas it is the same as the Web relevance score when $\gamma = 1$.

Experiments

Two kinds of the experiment were carried out using a Japanese Q&A site corpus, i.e., the 100 questions of "NTCIR-ACLIA2" (Mitamura et al. 2010), as the test questions. The "Yahoo! Chiebukuro" data were used as examples of a Q&A site corpus for calculation of mutual information and for training of the translation probability. The "Yahoo! Chiebukuro" data is distributed to researchers from the National Institute of Informatics based on a contract with the Yahoo Japan Corporation (National Institute of Informatics 2009) . "Yahoo! Chiebukuro" is the largest knowledge retrieval service in Japan, and the Yahoo Japan Corporation has been providing this service since April 2004. Their aim is to connect people who want to question and those who want to answer, and the sharing of wisdom and knowledge among the participants. The National Institute of Informatics provides data consisting of 3.11 million questions and

13.47 million answers (total text size of 1.6 billion characters) submitted between April 2004 and October 2005 out of about 10 million questions and about 35 million answers currently stored. The 100 questions of "NTCIR-ACLIA2" is included in NTCIR-8 ACLIA test collection (National Institute of Informatics 2012). This test collection includes 100 Japanese topics of Mainichi News Paper, which consists of 377,941 documents between years 2004 and 2005. It can be used for experiments of Complex Question Answering.

Morphological analysis was only carried out in the question analysis module although some works such as (Oda and Akiba 2009) and (Mizuno et al. 2007) classified question types there. ChaSen (Kyoto University and NTT 2013) was used as a morphological analyzer and the Yahoo! API (Yahoo Japan Corporation 2013) was used as a search engine. A candidate answer is always a sentence since we did not classify the question type. Web documents were retrieved for a query of all the question's content words with or without query expansion and were used as the source of the candidate answers. Thus, we did not have tagged answers.

Experiments of query expansion

The experiments were carried out as follows for each question. First, words were chosen from answers of a Q&A corpus as candidates for the query expansion. Here, each single word was chosen for every combination of two words in question for the system of the proposed method. By contrast, each word was chosen for every word in question for the system of the original method. Next, the top three words at most are chosen in the order of mutual information as the words to be added for the query expansion. Finally, the candidate answers were retrieved and evaluated.

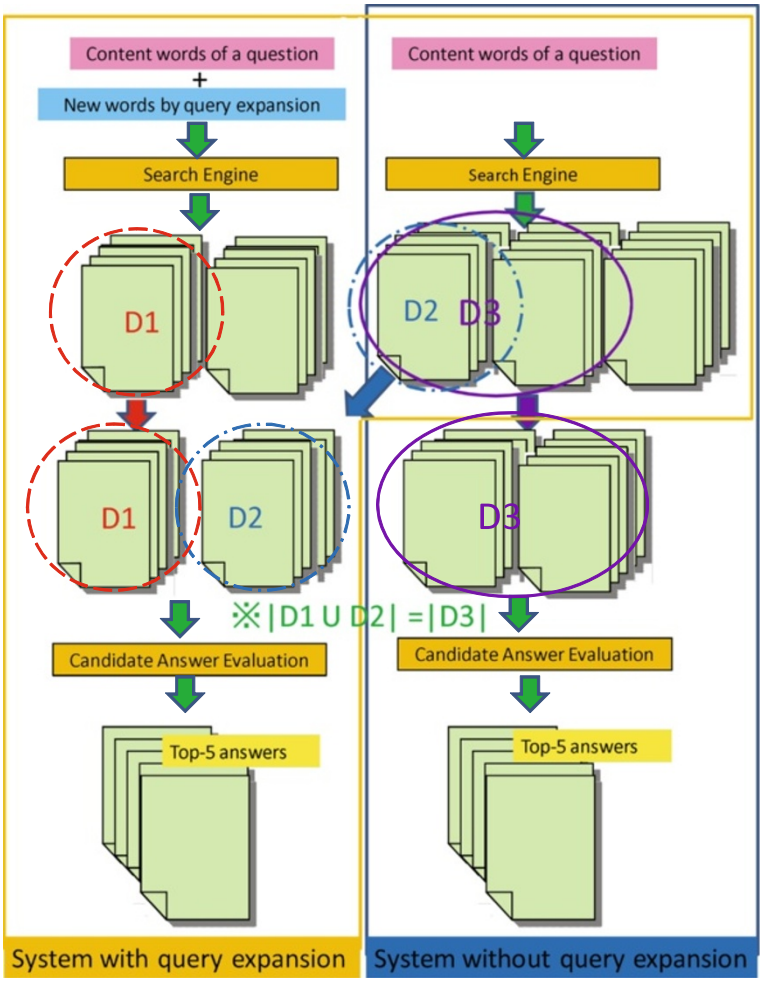

Figure 1 Outlines of Web document retrieval and candidate answer evaluation with and without query expansion. This figure shows the outlines of the Web document retrieval and the candidate answer evaluation with and without query expansion. *D*1 ∪ *D*2 and *D*3 each represents the candidate answers with and without query expansion. The number of the documents of *D*1 ∪ *D*2 was equalized to that of *D*3 for the fair comparison.

Figure 1 shows the outlines of the Web document retrieval and the candidate answer evaluation of the systems with and without query expansion. The documents were retrieved from the Web two times for the system with the query expansion: using all content words and using all content words and all new words for query expansion. The candidate answers were collected from the two sets of document retrieved by the system. On the other hand, the documents were retrieved from the Web only once for the system without query expansion: using all content words. The candidate answers were collected from them. $D1$ is a subset of documents retrieved by a query with the query expansion and $D2$ and $D3$ are subsets of documents retrieved by a query without query expansion in Figure 1. $D1 \cup D2$ and $D3$ each represents the candidate answers with and without query expansion. The number of the documents of $D1 \cup D2$ was equalized to that of $D3$ for the fair comparison; we set it to 80 documents. The score proposed by Ishioroshi et al. (2009) (the score of equ. (7)) was used for the candidate answer evaluation. The weighting parameter γ is set to 0.5. Unigrams were used as the feature of the writing style.

Experiments of candidate answer evaluation

GIZA++ (Casacuberta and Vidal 2007), which is the implementation of IBM-Model1, was used as a learning tool for the translation probability. The number of iterations of EM-algorithm was set to five times. The examples of a Q&A site corpus whose question or answer contains more than 60 words were preliminarily cut off because they negatively affected the learning of word alignment; they contained too many words. Moreover, the examples of a Q&A site corpus whose number of the words in the question is more than five times as many as that in the answer were cut off and vice versa for the same reason. As a result, 1,092,144 examples in the "Yahoo! Chiebukuro" data were used as the training data of GIZA++.

Fifty Web documents retrieved for a query without query expansion were chosen as the candidate answers and were evaluated by the proposed or original formula of the answer evaluation.

The bigrams normalized by the number of words were used for $P(A)$.

Results

Each candidate answer retrieved from Web documents was evaluated in the answer evaluation module and the QA system output the top-5 answers. The outputs of the system were checked manually. The top-5 accuracies and the MRR (Mean Reciprocal Rank) of the QA system were evaluated. The answer the system output is correct if it is in the top-5 answers when the top-5 accuracy is calculated. The top-5 accuracy is formulated as follows:

$$top-5\_accuracy = \frac{answered\_question}{N}, \tag{14}$$

where *answered_question* is the number of the question where the system output the correct answer in the top-5 answers. MRR is formulated as follows:

$$MRR = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{rank(i)}, \tag{15}$$

where $rank(i)$ represents the best rank of the correct answer of the ith question. MRR takes into consideration the rank of the output whereas the top-5 accuracy does not.

Results of query expansion

Table 1 shows the top-5 accuracies and MRR of the experiments of the query expansion. The original method in the table represents the method of Berger et al. (2000), where words to be added are chosen based on mutual information between a word from each question and another word from its answer. This table shows the system with the proposed method outperformed the system without query expansion and the system with the method of Berger et al. (2000). It also showed that the system with the method of Berger et al. (2000) is inferior to the system without query expansion. We think this is because the large corpus we used caused the semantic drift. Thus, we think the method of Berger et al. (2000) is unsuitable for the open-domain QA.

On the other hand, the proposed method can choose words to be added for the query expansion without the semantic drift, because it considers the co-occurrence of not only one word but also two words from each question and another word from an answer. The difference between the original method and the proposed method was significant though the difference between the system without query expansion and the proposed method was not, according to a Wilcoxon signed-rank test. The significance level was 0.05.

Table 1 Results of experiments of query expansion

	Without query expansion	Original method	Proposed method
Accuracy	0.420	0.400	0.450
MRR	0.262	0.233	0.273

This table summarizes the top-5 accuracies and MRR of the systems for the experiments of the query expansion. Original method in the table represents the method proposed by Berger et al. (2000), where the words to be added are chosen based on mutual information between a word from a question and another word in its answer. This table indicates that the system with the proposed method outperformed the two systems: the system without query expansion and the system with the method proposed by Berger et al. (2000).

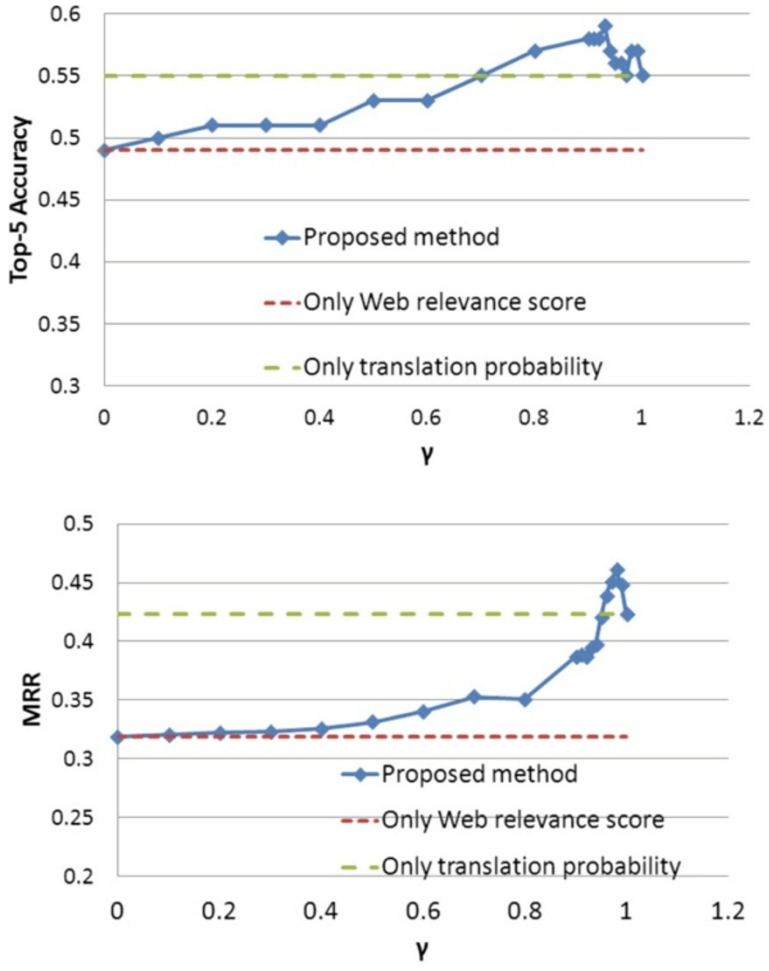

Figure 2 Performance of proposed method. This figure shows the top-5 accuracies and MRR of the experiments of the candidate answer evaluation when the value of γ changed from 0 to 1. The top-5 accuracy was maximized to 0.59 when $\gamma = 0.93$ and the MRR was maximized to 0.461 when γ =0.98.

Results of candidate answer evaluation

Figure 2 shows the top-5 accuracies and MRR of the experiments of the candidate answer evaluation when the value of γ changed from 0 to 1. Table 2 lists the performances of the original methods and the proposed method. Table 2 shows that the top-5 accuracy was maximized to 0.59 when γ = 0.93. In addition, the MRR was maximized to 0.461 when γ =0.98. As for the MRR, the proposed method was significantly better than the original methods according to a Wilcoxon signed-rank test. The significance level was 0.05.

Table 2 Results of experiments of answer candidate evaluation

	Top-5 accuracy	MRR
Only Web relevanve ($\gamma = 1$)	0.55	0.423
Only translation probability ($\gamma = 0$)	0.49	0.318
Proposed method ($\gamma = 0.93$)	0.59	0.395
Proposed method ($\gamma = 0.98$)	0.57	0.461

This table summarizes the top-5 accuracies and MRR of the systems for the experiments of the candidate answer evaluation. As for MRR, the proposed method was significantly better than the original methods according to a Wilcoxon signed-rank test.

Discussion

We will show examples of the results and discuss them in this section.

Query expansion

The examples A, B, and C show the examples from the experimental results.

> **Example A** Some examples where the new found word was the answer to the question could be found.

Question
"ソフトバンクとヤフーはどんな関係にありますか?"
"(What are the connections between softbank and yahoo?)"
Query expansion via the original method
{ソフトバンク (softbank) → ホークス (hawks)} and {ヤフー (yahoo) → メール (mail)}
Query expansion via the proposed method
{ソフトバンク (softbank) and ヤフー (yahoo) → 子会社 (subsidiary)}

According to Example A, we can see that the direct answer to the question was selected as a word for the query expansion via the proposed method. Even if the word *subsidiary* is not the direct answer, it is suitable for the query expansion because it has close connections with *softbank* and *hawk*.

Example B Some examples where the new found word was a clue to the question were also found.

Question
"インドとパキスタンの和平交渉に関する出来事を挙げて下さい."
"(Let me know events related to the peace plan of India and Pakistan?)"
Query expansion via the original method
インド (India) → カレー (curry)} and {パキスタン (Pakistan) → イスラム (Islamic)}
Query expansion via the proposed method
{インド (India) and パキスタン (Pakistan) → カシミール (Kashmir)}

Example B is a QA where the system cannot answer in a word; it is a non-factoid question. *Kashmir* is an important word because it is area that is close to India and Pakistan. On the other hand, *curry*, the word that is irrelevant to the question, was chosen via the original method. These words would cause the semantic drift, which sometimes makes it difficult to find documents that are relevant to the question. These words were frequently chosen via the original method, which decreased the performance of the system. We think that these cases did not happen in the experiments by Berger et al. (2000) because they used relatively small and domain specific corpora,

On the contrary, the proposed method where the system chooses the words that maximize mutual information between two words from a question and one word from its answer chose these words less frequently than the original method. It enabled better document retrieval for the QA that is not domain specific.

Example C These were some examples where the new found words were irrelevant to the question.

Question
"2004 年に原油価格が高騰したのはなぜですか？"
"(Why did the price of crude increase in year 2004?)"
Query expansion via the original method
{高騰 (increase) → 石油 (oil)} and {年 (year, age) → 結婚 (marriage)}
Query expansion via the proposed method
{高騰 (increase) and 年 (year, age) → 選手 (player or athlete)}
{する (does) and 年 (year, age) → 結婚 (marriage)}

Words that are irrelevant to the question were chosen for the query expansion even via the proposed method in Example C. There were many cases like them when general words were used for the calculation of mutual information. Therefore, we think that the words to calculate mutual information should be carefully selected in the future.

Candidate answer evaluation

Web relevance score

We will discuss about how scores of the topic relevance from the Web contributed the results. Examples D and E have examples of the Web relevance score for factoid and non-factoid questions, respectively. Web relevance scores of the words in answers are shown in brackets. Those of the words in questions were omitted.

Example D The relevant words could be obtained via the Web relevance score for some factoid questions.

Question
"2008 年のオリンピックはどこで開催されますか."
"(Where does the Olympic hold in 2008?)"
Example of answer
"北京" "(Beijing)"
The relevant words
北京, Beijing (0.67), 地, place (0.29), 大会, convention (0.29), 11 (0.24), 2011 (0.24), 12 (0.23), 回, times (0.23), 日本, Japan (0.23), 東京, Tokyo (0.22), 代表, represent (0.21), 競技, game (0.2), 20 (0.2), 夏季, summer (0.19), 冬季, winter (0.19), 中国, China (0.19)···

Direct answers could be obtained when the question was factoid as shown in Example D. We could particularly obtain *Beijing*, which was related to both *2008* and *Olympic*, although we could hardly obtain these words via the method using only translation probability that can only take into consideration one word at a time.

Example E The relevant words could be obtained via the Web relevance score for some factoid questions.

Question
"ＥＳ細胞とはどのようなものですか."
"(What are ES cells?)"
Example of answer
"体の様々な組織の細胞になる能力がある,
人の胚性幹細胞（ES 細胞）を,
大量に培養し効率よく
大脳の細胞にすることに,
理化学研究所発生・
再生科学総合研究センター
（神戸市）の笹井芳樹・
グループディレクターらの
研究チームが成功した．"
"(The research team of Center of Developmental Biology, Institute of Physical and Chemical Research in Kobe, whose leader is Yoshiki Sasai, succeeded in largely culturing embryo-stem cells (ES cells), which have ability to be any types of cells in various tissues in body, and effectively changing them into cerebral cells.)"
The relevant words
性, sex (0.29), 研究, research (0.26), 幹, stem (0.24), ヒト, human (0.23), 胚, embryo (0.2), 分化, differentiation (0.18), マウス, mouse, mice (0.18), 再生, remodeling, regeneration (0.16), 組織, tissue (0.15), 医療, medical (0.13), 的, like (0.13), 体, body (0.13), 科学, science (0.12), 培養, culture (0.11), iPS (0.11),· · ·

Direct answers to the question could not be obtained when the question was non-factoid. However, the words that are related to the question could be obtained. The suitable answers that include the relevant words could be also obtained as shown in Example E. However, words that frequently appear in many documents could not be distinguished from those that co-occur with content words in the question using mutual information. Thus, we think that the selection of these words using IDF will be able to be tried in the future.

Translation probability

We will discuss about how the translation probability contributed the results.

Table 3 has examples of the top-5 words that maximize $P(q|a)$, which is the translation probability from a word a in an answer to a word q in a question when a is given. The English words and the numbers in brackets are the English translations and the translation probabilities, respectively. For example, when "医療" (medical care) was given as a word in an answer, it tended to be translated into "医療" (medical care), "病院" (hospital), "費" (fare), "入院" (medical admission), and "手術" (operation) in its question. This indicates that "医療" (medical care) tends to appear in the answer when these words appear in its question. The functions of Japanese words are shown when the English words are written in upper case.

Table 3 firstly shows words in answers are likely to be translated into themselves in their questions. This indicates that words in questions tend to appear in their answers. Next, the table shows words in answers are likely to be translated into their relevant words and synonyms as shown in the case where (1) "入院" (medical admission) and "手術" (operation) for "医療" (medical care), and (11) "首相" (prime minister) for "総理" (prime minister) are listed in the table. This indicates that relevant words and synonyms of words in question tend to appear in their answers.

The properties of the relevant words and the synonyms that were obtained using the translation probability are different from those obtained from 100 Web documents because they were from approximately one million examples of a Q&A site corpus. Therefore, we think that the performance of the QA system improved because the Web relevance score and the translation probability complemented one another.

We expected that (13) "だから" (because), (14) "から" (because, from), and (15) "ため" (because, for) were likely to co-occur with "なぜ" (why) or "どうして" (why), which often appeared in questions, because they often appeared in answers of QA, but they did not. We think that this is because the particles like "から" (because, from) and "ため" (because, for) are ambiguous. Soricut and Brill (2006), who used an English Q&A corpus for learning, reported that "because" tended to be translated into "why". We think that the method worked well because the English word "because" was less ambiguous than Japanese words like "から" (because, from) and "ため" (because, for).

However, (16) "理由" (reason) , which is also likely to appear in answers to why-type questions, could be leaned as the word that tended to be translated into "なぜ" (why). This indicates that learning with the translation probability could be able to partially evaluate the writing style.

In addition, the words that appeared few times tended to be learned not correctly. For example, (12) "ナズナ" (shepherd's-purse) were hardly translated into relevant words because it appeared only twice in the Q&A site corpus. Moreover, some unsuitable words were chosen because the translation probabilities only depended on the Q&A site corpus. The "Yahoo! Chiebukuro" data are examples of Q&A site submitted from April 1st 2004 to October 31th 2005. Therefore, "小泉" (Koizumi), who was

Table 3 Examples of translation probability

Index	Given word	1st	2nd	3rd	4th	5th
(1)	医療	医療 (.064)	病院 (.057)	費 (.037)	入院 (.026)	手術 (.024)
	Medical care	Medical care	Hospital	Fare	Admission	Operation
(2)	訴訟	訴訟(.097)	裁判 (.032)	訴える (.021)	弁護士 (.016)	権 (.016)
	Lawsuit	Lawsuit	Judgment	Sue over	Advocate	Right
(3)	塩水	塩水 (.081)	水 (.034)	塩味 (.025)	方法 (.023)	殻 (.022)
	Salt water	Salt water	Water	Tastes salty	Method	Shell
(4)	地形	地形 (.033)	横浜 (.020)	たび (.019)	台風 (.015)	地理 (.015)
	Landform	Landform	Yokohama	Times	Typhoon	Geography
(5)	海峡	海峡 (.115)	島 (.0373)	世界 (.024)	竹島 (.024)	東北 (.021)
	Channel	Channel	Island	World	Takeshima	Tohoku
(6)	逮捕	逮捕 (.249)	れる (.060)	捕まる (.031)	する (.028)	た (.023)
	Arrestment	Arrestment	PASSIVE	Get caught	Do	PAST
(7)	細胞	細胞 (.132)	? (.031)	です (.030)	なぜ (.029)	人間 (.026)
	Cell	Cell	?	PREDICATION	Why	Human
(8)	情報	情報 (.146)	の (.060)	です (.031)	する (.031)	が (.030)
	Information	Information	Of	PREDICATION	Do	AGENT
(9)	技術	技術 (.134)	の (.065)	は (.040)	です (.039)	?(.038)
	Technology	Technology	Of	TOPIC MARKER	PREDICATION	?
(10)	大統領	大統領 (.298)	アメリカ (.098)	ブッシュ (.039)	の (.025)	?(.021)
	President	President	America	Bush	Of	?
(11)	総理	総理 (.222)	小泉 (.138)	大臣 (.047)	さん (.040)	首相 (.030)
	Prime minister	Prime minister	Koizumi	Minister	Mr.	Prime minister
(12)	ナズナ	七草 (.191)	?(.170)	って (.143)	の (.079)	か (.045)
	Shepherd's-purse	Seven herbs	?	As for	Of	QUESTION
(13)	だから	?(.064)	の (.063)	です (.052)	は (.049)	か (.046)
	Because	?	Of	PREDICATION	TOPIC MARKER	QUESTION
(14)	から	の (.073)	です (.058)	? (.056)	か (.050)	が (.046)
	Because	Of	PREDICATION	?	QUESTION	AGENT
(15)	ため	の (.088)	です (.061)	? (.055)	か (.052)	ため (.048)
	For	Of	PREDICATION	?	QUESTION	For
(16)	理由	理由 (0.17)	なぜ (0.04)	が (0.04)	の (0.04)	か (0.03)
	Reason	Reason	Why	AGENT	Of	QUESTION

This table has examples of the top-5 words that maximize $P(q|a)$, which is the translation probability from a word a in an answer to a word q in a question when a is given. The English words and the numbers in brackets are the English translations and the translation probabilities, respectively. The functions of Japanese words are shown when the English words are written in upper case.

the prime minister at that time, and "Bush", who was the president of USA at that time, were chosen as the words likely to be translated from "総理" (prime minister) and "大統領" (president), respectively.

Conclusion

Question Answering (QA) is a task of answering natural language questions with adequate sentences. It includes the relevant document retrieval and candidate answer evaluation modules. This paper proposed two methods to improve the performance of the QA system using a Q&A site corpus. The first method is for the query expansion in the relevant document retrieval module. We proposed modification of measure of mutual information for the query expansion; we calculate it between two words in each question and a word in its answer in the Q&A site corpus not to choose the words that are not suitable. The second method is for the candidate answer

evaluation module. We proposed the method to evaluate candidate answers using existing two methods, i.e., the Web relevance score and the translation probability. We showed that the proposed method evaluated the candidate answers more effectively than the original methods. The experiments were carried out using a Japanese Q&A site corpus. They revealed that the first method was significantly better than the original method when the accuracies and MRR were compared. They also showed that the second method was significantly better than the original methods when the MRR were compared.

Endnotes

[a] Berger et al. (2000) used Usenet FAQ documents and customer service call-center dialogues from a large retail company.

[b] We got this word because we had a baseball team named softbank hawks in Japan.

[c] $P(q_j|a_i)$ was summed from 1 to $l+1$ because each question word had exactly one connection to either a single answer word or empty.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

KK is the first author of this journal and one of the original conference papers of this journal paper (Komiya et al. 2013). She was an assistant professor of the laboratory that YA belonged to and she advised him on the key idea of the first method for the relevant document retrieval module. YA is the first author of another original conference paper of this journal paper (Abe et al. 2013). He carried out experiments of this journal for his master thesis. This paper is based on his thesis written in Japanese. HM advised YA on the second method for the candidate answer evaluation module. YK was a supervisor of YA and professor of the laboratory that YA belonged to. He also advised YA and KK on all the papers they wrote relevant to this journal paper. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank Yahoo Japan Corporation and the National Institute of Informatics which provide us the "Yahoo! Chiebukuro" data. This work was partially supported by JSPS KAKENHI Grant Number 24700138.

Author details

[1] Institute of Engineering, Tokyo University of Agriculture and Technology, 2-24-16 Nakacho-Kotanei, Tokyo, 184-8588, Japan. [2] Tokyo Institute of Technology, 4259 Nagatsuta, Midori-ku, Yokohama, Kanagawa, 226-8503, Japan.

Received: 10 November 2012 Accepted: 26 July 2013
Published: 22 August 2013

References

Abe Y, Morita H, Komiya K, Kotani Y (2013) Question Answering System Using Web Relevance Score and Translation Probability In: Proceedings of JCKBSE 2012, Frontiers in AI and Applications, Vol.180. IOS Press Rhodos, Greece, pp 11–15

Berger A, Caruana R, Cohn D, Freitag D, Mittal V (2000) Bridging the lexical chasm: Statistical approaches to answer-finding In: Proceedings of SIGIR. ACM, Athens, pp 192–199

Brown PF, Della Pietra SA, Della Pietra VJ, Mercer RL (1993) The mathematics of statistical machine translation: parameter estimation" computational linguistics. J Inf Retrieval - Spec Issue Web Inf Retrieval 19(2): 263–311

Casacuberta F, Vidal E (2007) GIZA++: Training of statistical translation models. http://web.iti.upv.es/~evidal/students/doct/sht/transp/OLD/giza4p.pdf

Derczynski L, Wang J, Gaizauskas R, Greenwood MA (2008) A data drive approach to query expansion in question answering In: Proceedings of Coling 2008. Coling 2008 Organizing Committee, Manchester, pp 4–41

Higashinaka R, Isozaki H (2007) NTT' s question answering system for NTCIR-6 QAC-4 In: Working Notes of the NTCIR Workshop Meeting (NTCIR). National Institute of Informatics, Tokyo, pp 460–463

Higashinaka, R, Isozaki H (2008) Corpus-based question answering for why-questions In: Proceedings of IJCNLP. Asian Federation of Natural Language Processing, Hyderabad, pp 418–425

Ishioroshi M, Satol M, Mori T (2009) A web question-answering method for any class of non-factoid questions. J JSAI 24(4): 339–350

Isozaki H, Higashinaka R (2008) Web NAZEQA, a web-based why-question answering system In: Proceedings of the, IPSJ Kansai conference. Information Processing Society of Japan, Kyoto, pp 143–146

Komiya K, Abe Y, Kotani Y (2013) Query Expansion using Mutual Information between Words in Question and Its Answer In: Proceedings of ITI. CIT, Cavtat / Dubrovnik, pp 289–294

Kyoto University, NTT (2013) MeCab: Yet another part-of-speech and morphological analyzer. http://mecab.sourceforge.net/

Lao N, Shima H, Mitamura T, Nyberg E (2008) Query expansion and machine translation for robust cross-lingual information retrieval In: Proceedings of NTCIR-7 Workshop Meeting. National Institute of Informatics, Tokyo, pp 140–147

Lin MC, Li MX, Hsu CC, Wu SH (2010) Query expansion from Wikipedia and topic web crawler on CLIR In: Proceedings of, NTCIR-8 Workshop Meeting. National Institute of Informatics, Tokyo, pp 101–106

Liu M, Zhou B, Qi L, Zhang Z (2010) Wikipedia article content based query expansion in IR4QA system In: Proceedings of NTCIR-8 Workshop Meeting. National Institute of Informatics, Tokyo, pp 136–139

Mitamura T, Shima H, Sakai T, Kando N, Mori T, Takeda K, Lin CY, Song R (2010) Overview of the NTCIR-8 ACLIA Tasks: Advanced Cross-Lingual Information Access In: Proceedings of 8th NTCIR Workshop Meeting. National Institute of Informatics, Tokyo, pp 15–24

Mizuno J, Akiba T, Fujii A, Itou K (2007) Non-factoid question answering experiments at NTCIR-6: Towards answer type detection for realworld questions In: Proceedings of 6th NTCIR Workshop Meeting. National Institute of Informatics, Tokyo, pp 487–492

Mori T, Sato M, Ishioroshi M, Nishikawa Y, Nakano S, Kimura K (2007) A monolithic approach and a type-by-type approach for non-factoid question-answering In: Proceedings of 6th NTCIR Workshop Meeting. National Institute of Informatics, Tokyo, pp 469–476

National Institute of Informatics (2009) Distribution of "Yahoo! Chiebukuro" data. http://www.nii.ac.jp/cscenter/idr/yahoo/tdc/chiebukuro_e.html

National Institute of, Informatics (2012) NTCIR Project NTCIR-8 ACLIA. http://research.nii.ac.jp/ntcir/permission/ntcir-8/perm-en-ACLIA.html

Oda T, Akiba T (2009) Improving type identification method for non-factoid question answering. J Forum Inf Technol 8(2): 265–268

Saggion H, Gaizauskas R (2004) Mining on-line sources for definition knowledge In: Proceedings of FLAIRS. AAAI Press, Miami Beach, pp 45–52

Soricut R, Brill E (2006) Automatic question answering using the web: beyond the factoid. J Inf Retrieval - Spec Issue Web Retrieval 9: 191–206

Yahoo Japan Corporation (2013) Yahoo! Japan Developer Network. http://developer.yahoo.co.jp

doi:10.1186/2193-1801-2-396
Cite this article as: Komiya *et al.*: Question answering system using Q & A site corpus Query expansion and answer candidate evaluation. *SpringerPlus* 2013 **2**:396.