Introduction of a Probabilistic Language Model to Non-Factoid Question-Answering Using Example Q&A Pairs

Kyosuke Yoshida, Taro Ueda, Madoka Ishioroshi, Hideyuki Shibuki, and Tatsunori Mori
Graduate School of Environment and Information Sciences, Yokohama National University
79-7 Tokiwadai, Hodogaya-ku,Yokohama 240-8501, Japan
{kyoshida, kks, ishioroshi, shib, mori}@forest.eis.ynu.ac.jp

Abstract

In this paper, we propose a method which utilizes a probabilistic language model in non-factoid type question-answering system in order to improve its accuracy. The model is a mixture probabilistic language model of part-of-speech and surface expressions. We introduced the model into two sub-processes which calculate similarity of texts in terms of writing style. The first process collects example questions similar to a submitted question. The second one measures similarity between an answer candidate and example answers paired with the collected example questions. Experimental results showed that the accuracy of the system was improved by introducing the proposed method.

1 Introduction

In recent years, the amount of data available on the Web is increasing by growing computer performance and network traffic. Therefore, technologies that give us access to neccessary information in the large amount of data are required. One of such technologies is question-answering (QA), which is to extract an answer for a question written in natural language from source documents. In general, QA systems are categorized into the following two types: factoid and non-factoid(Fukumoto, 2007). We focus on the non-factoid type QA in this paper. Table 1 shows some typical types of non-factoid questions. The appropriateness of the answer candidates is often estimated on the basis of following two measures (Han *et al.*, 2006).

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

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. Although these two measures depend on each other to some extent, we assume that they are independent in this study.

Non-factoid type QA systems are categorized into the following two types according to how to handle Measure 2. The first type classifies submitted questions into several predefined question types such as definition-type, why-type, how-type, and so on, in order to separately handle each type of questions by different methodologies. Han *et al.* (2006) calculated the above-mentioned two measures for definition-type questions based on probabilistic models built from corpora. The model for Measure 1 is calculated from retrieved documents. The model for Measure 2 is calculated from a corpus of definitions. However, this type of systems has some difficulties as follows. Since 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 different classes.

The second type of systems handles submitted questions based on a unified framework without question classification. Mizuno *et al.* (2009) proposed a method that is able to calculate Measure 2 without classification of questions. Using example Q&A pairs from a Q&A community site, it learns a binary classifier that judges whether or not the class

26th Pacific Asia Conference on Language,Information and Computation pages 361–370

Table 1: Typical types of non-factoid questions

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

of a given answer candidate is consistent with the class of a submitted question. By using this classifier, Measure 2 is realized without question classification. Soricut et al. (2006) also proposed a system without question classification. They introduced a statistical translation model between questions and the corresponding answers in order to bridge the lexical gap between the questions and the answers. A set of example Q&A pairs from FAQ sites on the Web is used for the estimation of the model.

In these methods, the length of answers should be predetermined. The length of answers cannot be changed dynamically and is necessary to be estimate from the length of the question.

Therefore, Mori *et al.* (2008) proposed a method of the second type approach that is able to adaptively determine the length of an answer candidate according to a submitted question. They use example Q&A pairs on a Q&A community site in order to find appropriate writing styles to answers for submitted questions. They utilize simple n-gram model as features to retrieve example questions similar to a submitted question in terms of writing style and to find appropriate writing styles to answer. However, the simple n-gram model is not appropriate to model dependency among words that appear in the distant positions because it only captures linguistic phenomena that appear within the n-words window. Therefore, sometimes the selection of example questions is not carried out correctly. There exist some incorrectly retrieved example questions that are not similar to the whole submitted question in terms of writing style, while those n-grams happen to be very similar to the n-grams of the question. Their method of scoring answer candidate is based on a naive frequency model of word 2-grams as feature expressions. Therefore, ungrammatical sentences, which often appear in Web documents and are not suitable to answer candidates, happen to have high scores when they have the feature expressions. It decrease the accuracy of the system.

In this paper, we employ the method of Mori *et al.* (2008) as a baseline method. We introduce a probabilistic language model to the baseline method in order to solve the above problems and improve the method in terms of accuracy.

Our method has the following three feature parts. Firstly, a probabilistic language model is used to retrieve examples similar to an submitted question. Secondly, another probabilistic language model is constructed from the retrieved example answers, which is used to measure the appropriateness of answer candidates for submitted questions. Finally, the answer candidates are clustered into several groups, and the candidates that have unsuitable writing styles as answers for the submitted question are removed.

The rest of this paper is organized as follows. In Section 2, we explain the related works. In Section 3, we explain the outline of the baseline method. In Section 4, we discuss the problems of the baseline method. In Section 5, we describe the detail of the proposed method. In Section 6, we conduct examinations of our QA system, and discuss the results. In Section 7, we provide our conclusion.

2 Related Works

The methods which utilize probabilistic language model have been developed including followings.

Takahashi *et al.* (2010) combine several types of language models in order to retrieve questions similar to users' queries from a Q&A archive of a Q&A community site. In order to examine the mixture ratio of the language models, they investigated the following two cases: 1) the ratio is fixed for all Q&A pairs, and 2) the ratio adaptively varies according to Q&A pairs. They showed that the performance is improved in both of the cases. The purpose of this study is different from ours because we retrieve example questions similar to submitted question in terms of writing styles while they retrieve questions similar to submitted questions in terms of content.

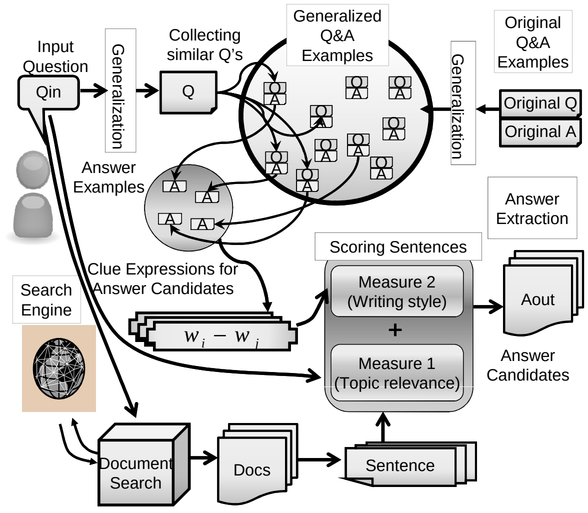

Figure 1: Outline of the baseline system

Heie *et al.* (2012) proposed a method to obtain answers by calculating the relation between the submitted question Q and an answer candidate A in terms of probability. They supposed that the probability of having the answer A depends on two sets of feateures, W and X, as $P(A|Q) = P(A|W, X)$. The set of feateures W $(= w_1, \ldots, w_{|W|})$ denotes feature expressions that indicate "type of question" , e.g. "when", "why", "how". 2,522 words are obtained from TREC question set as the candidate of W. X $(= x_1, \ldots, x_{|X|})$ denotes a set of features comprising the "information-bearing" words of submitted quesions, e.g. what the question is actually about and what it refers to. They used $P(A|X)$ as a retrieval model and $P(W|A)$ as a filter model.

Although two above-mentioned studies do not explicitly handle the questions in a question-type-by-question-type manner, they explicitly use surface expressions. On the other hand, our method take account of not only surface expressions but also their part-of-speech tags as their abstractions. In order to take account of writing styles, we utilize a mixture probabilistic language model in terms of part-of-speech tags and surface expressions.

3 Baseline Method

In this section, we describe the baseline method according to Mori *et al.* (2008). Figure 1 shows the outline of the baseline QA system.

3.1 Extracting Keywords from a Question and Obtaining Their Related Words

From a question submitted by a user(a submitted question, hereafter), 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 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 the obtained snippets.

3. The weight value $T(w_j)$ defined as the following equation is calculated for each word w_j in snippets:

$$T(w_j) = \max_i \frac{freq(w_j, i)}{n_i} \qquad (1)$$

   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:

$$T(k) = \max_j T(w_j) \qquad (2)$$

3.2 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 baseline method retrieves example Q&A pairs whose questions are similar to the submitted question from the viewpoint of writing style. Mori *et al.* (2008) adopted the word 7-gram whose center word is an interrogative as the core part of a given question, because it represents enough context to determin the class of question. Thorefore, they defined the similarity between two questions as the similarity between the word 7-grams extracted from the questions. According to the similarity, N-best example Q&A pairs are obtained by using an ordinary information retrieval technique.

3.3 Extracting Clue Expressions from Example Answers

In this stage, clue expressions are extracted from the answers in the example Q&A pairs obtained in the stage described in Section 3.2. A 2-gram was adopted as a clue expression unit because it is the smallest unit that can represent relations between words. It is assumed 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 retrieved Q&A pairs.

As the measurement of the correlation, Mori *et al.* (2008) adopted the χ^2 value shown in Equation (3) for the following two kinds of events for the answers from the entire set of example Q&A pairs:

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

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

$$\chi^2(b) = \frac{n}{|A| \cdot |\bar{A}| \cdot |B(b)| \cdot |\bar{B}(b)|} \cdot (|A \cap B(b)| \cdot |\bar{A} \cap \bar{B}(b)| - |\bar{A} \cap B(b)| \cdot |A \cap \bar{B}(b)|)^2 \quad (3)$$

where n is the total number of example Q&A pairs. The more correlated two events are, the larger the value of $\chi^2(b)$ 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.

3.4 Extracting Answer Candidates

In this stage, by using the method in Section 3.3, it extracts a set of 2-grams as clue expressions from the example answers of the example Q&A pairs retrieved by the method in Section 3.2 and calculates the corresponding $\chi^2(b)$ value for each 2-gram b. The score of each sentence is calculated by using the following equation:

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

where l is the number of different words in the sentence S_i, m is the numer of different 2-grams in S_i, w_{ij} is the j-th word in sentence S_i, and b_{ik} is the k-th 2-gram in S_i. Since the terms $\sum_{j=1}^{l} T(w_{ij})$ and $\sum_{k=1}^{m} \sqrt{\chi^2(b_{ik})}$ in Equation (4) correspond to Measure 1 and Measure 2, respectively, the parametar γ is used to determine the mixture ratio of Measure 1 and Measure 2. The normalization term $\frac{1}{\log(1+|S_i|)}$ is inroduced 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 correlated with example answers). In order to reward longer sentences, the logarithm of sentence length is adopted.

4 Problems of Baseline Method

In the baseline method, the $\chi^2(b)$ value of a word 2-gram mentioned in Section 3.3 is used in order to extract clue expressions from example answers. This method uses only the frequency of word 2-grams for the purpose of calculation based on the $\chi^2(b)$ value. As a result, the word order and the contexts of clue expressions are ignored. In this method, example questions are retrieved according to the similarity between submitted question and example questions in terms of the 7-gram whose center word is an interrogative. However, the selection of example questions occasionally fails because some retrieved example questions are not similar to the submitted question in terms of the writing style of whole sentence in spite of high degree of similarity in terms of the 7-gram. The following is a submitted question and a wrongly-retrieved example Q&A pair which is not similar to the submitted question in terms of the writing sytle of whole sentence. The system handles Japanese texts. In the following example, the sentences written in italics are Japanese.

Question (submitted) : *BSE ga hito ni kansen suru to dou nari masu ka.*
(What happens for people when they are infected with BSE?)

Question (example) : *"Yuri no hana saku basho de" wo eigo ni suru to dou nari masu ka.*
(How do you say *"Yuri no hana saku basho de"* in English?)
Answer (example) : *"At the place where lilies bloom" desu.*
("At the place where lilies bloom" in English.)

In this example, the 7-grams are "kansen suru to dou nari masu ka" and "eigo ni suru to dou nari masu ka", and they are very similar to each other. However, they are very different from each other in terms of the writing style of the first half of sentences because the former is "noun (kansen) _ verb (suru) _ postposition (to)" and the latter is "noun (eigo)

_ postposition (ni) verb (suru) _ postposition (to)". They are also different from each other in terms of the topic of question because the former is "what the symptom is" and the latter is "translation in English of Japanese words". For these reasons, this example Q&A pair does not have a suitable writing style for the answer of the submitted question. The following is a retrieved example Q&A pair whose question part is similar to the submitted question, but whose answer part is not suitable as an answer to the submitted question in terms of writing style.

Question (submitted) : *Beikoku ga kyoutogiteisho wo hijun shi nai riyuu wa nan desu ka.*
(Why the U.S. government doesn't ratify Kyoto protocol?)

Question (example) : *Camping car wo katta riyuu wa nan desu ka.*
(Why did you purchase a camper?)
Answer (example) : *Trailer wo katte 7 nen ni nari masu. Katte yokatta desu.*
(It has been seven years since I purchased the camper. I'm glad I bought it.)

In this example, both questions ask a reason of an action, and the writing style of the example question is similar to one of the submitted question. However, the example answer is not an appropriate answer to the example question because it does not describe any reasons. Questions and answers in example QA pairs are not always consistent with each other, while the example answers correspondig to the example questions are the best answers in a QA community site. In this study, by resolving the above problems, we improve the baseline method in order for it to correctly retrieve the following question examples.

Question (submitted) : *Fog lamp wa nan no tame ni aru no desu ka.*
(What is a fog lamp for?)

Question (example) : *Mayuge wa nan no tame ni aru no desu ka.*
(What are eyebrows for?)
Answer (example) : *Ame ya ase ga me ni hairu no wo fusegu tame desu.*
(Because they prevent rains and sweat entering the eyes.)

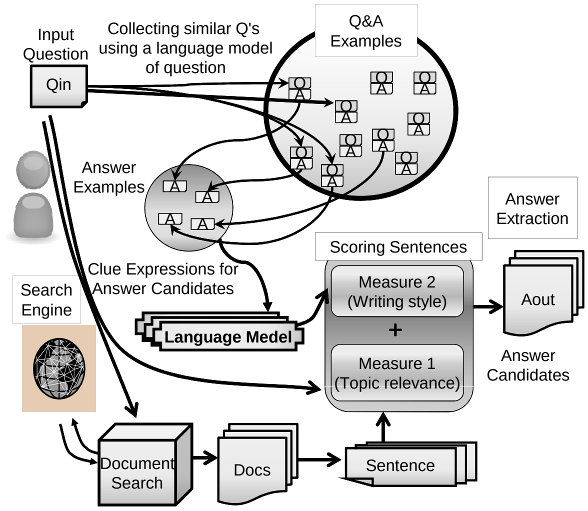

Figure 2: Outline of the proposed system

5 Proposed Method

In this study, we introduce probabilistic language models to following two processing steps. The first one is retrieving example questions similar to the submitted question mentioned in Section 3.2. The second one is extracting answer candidates mentioned in Section 3.3 and 3.4. In other words, we calculate Measure 2 by using the probabilistic language models instead of the original naive method. Our approach is expected to have the following three advantages.

- In the step of retrieving example questions, we can retrieve example questions that are more similar to the submitted question by using an appropriate probabilistic language model of question than example questions by using the baseline method because the probabilistic language model can take into account the effect of writing style in longer context, i.e., whole sentences.
- We can remove texts that include ungrammatical expressions and meaningless symbols from answer candidates by using an appropriate probabilistic language model of answer examples to extract answer candidates.
- We can remove example answers which have unsuitable writing style for the submitted question from example answers by using the language model of answer examples because we perform a clustering of example Q&A pairs by using skip 2-grams obtained from not only example questions but also example answers.

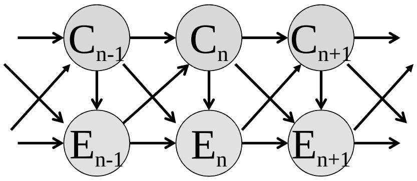

Figure 3: A mixture probabilistic language model in terms of part-of-speech tags and surface expressions

Figure 2 shows the outline of QA system we proposed.

5.1 Mixture Probabilistic Language Model in Terms of Part-of-Speech Tags and Surface Expressions

5.1.1 Outline

In the baseline method, 2-grams, which are used to extract clue expressions from example answers, are treated as the following two ways: 1) the following surface expressions are used as they are: the functional words (e.g. interrogatives particles and auxiliary verbs) and some predetermined content words that tend to express the focus of questions, 2) the other words are replaced with their part-of-speech tags in order to generalize them. However, it is unpredictable what words express the focuses of questions in the process of extracting clue expressions. Moreover, the words expressing focuses may vary according to question types and it is difficult to prepare a universal word list for any question type. In order to adaptively capture the adequate level of generalization of each word, i.e. adopting its surface expression as it is or its part-of-speech tags as generalization, we use a mixture probabilistic language model of part-of-speech tags and surface expressions. The model is shown in Figure 3. $P(E_1, E_2, \ldots, E_n)$, which is the probability of generating a sequence of surface expressions $E_1, E_2, \ldots, E_n$ as a sentence, may be estimated by using the mixture model as Equation (5).

$$\begin{aligned} P(E_1\,E_2 ... E_n) &\approx P(E_n|C_nE_{n-1}C_{n-1}) \\ &\cdot P(C_n|E_{n-1}C_{n-1}) \cdot P(E_1E_2...E_{n-1}) \\ &= \prod_{i=1}^{n}\{P(E_i|C_iE_{i-1}C_{i-1}) \cdot P(C_i|E_{i-1}C_{i-1})\} \end{aligned} \quad (5)$$

where C_i is the part-of-speech tag of E_i.

In order to adaptively determine the mixture ratio of surface expressions and their part-of-speech tags, we approximately estimate $P(E_1, E_2, \ldots, E_n)$ by a 2-gram model of words and their part-of-speech tags , which is obtained by a smoothing based on the deleted interpolation method.

5.1.2 Derivation of Generation Probability of a Given Sentence

We perform morphological analysis on a given sentence, divide the result of morphological analysis into a sequence of 2-grams and estimate a generation probability $P(E_1, E_2, \ldots, E_n)$ for the sequence by Equation (5).

5.2 Retrieving Example Questions Similar to the Submitted Question Using a Probabilistic Language Model

In this study, in order to retrieve example questions (along with their paired example answers) similar to the submitted question in terms of writing style, we obtain an optimal subset of example questions adaptively as follows: 1) generate subsets of example questions, 2) generate a language model from each subset, 3) calculate the generation probability of the submitted question for each language model, and 4) select the optimal subset, whose language model gives the highest probability to the submitted question. In other words, we retrieve subset of example questions which construct the best language model for the submitted question.

Ideally, the method can be implemented as the enumeration of all subsets in the above step 1), and the subsequent steps 2),3), and 4). Since, however, the corpus used in this study includes about 0.9 million Q&A pairs, the number of subsets explodes. Obviously it is not realistic to implement the method as above mentioned. Therefore, in order to shorten the processing time, we introduce an approximation based on the clustering according to the following procedure.

1. Determine the number of example questions which is retrieved finally. Let the number called "*target number*". In our experiments, we set it 500.

2. Retrive example questions (along with their paired answer examples) from a given Q&A corpus in descending order of similarity based on 7-gram mentioned in Section 3.2. In the baseline method, top-most example questions are simply employed as many as *target number* at this step. On the other hand, in the proposed method, we only utilize the 7-gram similarity as the first approximating to reduce the number of example questions. Let the number of exam-

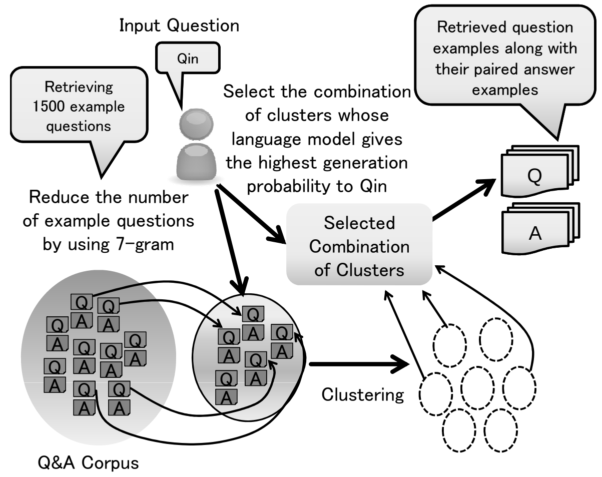
Input Question
Qin
Retrieving 1500 example questions
Select the combination of clusters whose language model gives the highest generation probability to Qin
Retrieved question examples along with their paired answer examples
Reduce the number of example questions by using 7-gram
Selected Combination of Clusters
Q
A
Clustering
Q&A Corpus

Figure 4: Retrieving question examples (along with answer examples) similar to an submitted question in terms of writing style

ple questions retrieved in this step three times of *target number*, in our experiment.

3. Apply a clustering algorithm to example questions extracted in the above step 2, and obtain several clusters.

4. Obtain combinations of clusters created in Step 3. Generate a probabilistic language model from the example questions in each combination of clusters. Calculate the generation probability of the submitted question for each model. Obtain the combination of clusters whose language model gives the highest probability to the submitted question.

The reason why we divide examples into some clusters is to shorten the processing time compared to calculating for all sebsets of example questions. The outline of this processing is shown in Figure 4.

5.2.1 Clustering Example Q&A Pairs

As described later in Section 5.3, we finally need to obtain example answers paired with the example questions that are similar to the submitted question. The clustering process described above is for not only example questions but also example answers, namely, for example Q&A pairs. In order to calculate similarity between sentences for clustering by taking account of word co-occurrence in distance positions of a sentence, we use word skip 2-grams as sentence features for clustering. A skip 2-gram is any pair of words in their sentence order. It may have some gaps between two words. Both question examples and answer examples are generalized for clustering (not for obtaining probabilistic language models) as follows: 1) the following surface expressions are used as they are: the functional words (e.g. interrogatives particles and auxiliary verbs) and some predetermined content words described below, 2) the other words are replaced with their part-of-speech tags. The predetermined context words includes a) words that tend to express the focus of question (e.g. “riyuu (reason)”, “houhou (method)”, “imi (meaning)”, “chigai (difference)”), and b) verbs and adjectives that frequently appear in corpus. As the words expressing the focuses of questions, we collect nouns X that frequently appear in the following contexts of corpus: “...X-wa nan-desuka (What is X of ...)”, “... X-wo oshiete (Tell me X of ...)”, and so on.

There are, at least, following three choices for similarity calculation when we cluster example questions and answers into some clusters.

Similarity 1

Similarity between Q&A pairs in terms of skip 2-grams. We take account of both the question part and the answer part of a Q&A pair simultaneously.

Similarity 2

Similarity between example questions only in terms of skip 2-grams.

Similarity 3

Similarity between example answers only in terms of skip 2-grams.

In the calculation of Similarity 1, we calculate the similarity of the question parts and that of the answer parts separately, then mix the values into one similarity, because the feature expressions from the answer parts should be treated independent of those of the question parts, and vice versa. As the clustering algorithm, we employed the k-means method.

5.2.2 Obtain the Optimal Combinations of Clusters

We employed a simple hill climbing method to retrieve the optimal combination of clusters whose language model of question parts gives the maximal generation probability to the submitted question. We use Equation (5) to calculate the generation probability and the combination is greedily searched through the following steps.

1. Let the cluster set CL be the given cluster set, and let the candidate set CA be an empty set.

2. In *CL*, find the cluster whose language model of question parts gives the maximum probability

Table 2: Case of use of Similarity 1 (using both question part and answer part) in clustering examples Q&A pairs

	Proposed method ($\gamma = 0.7$)		Proposed method ($\gamma = 0.8$)		Proposed method ($\gamma = 0.9$)		Baseline ($\gamma = 0.5$)	
Type of Question	MRR	Number of Correct Response	MRR	Number of Correct Response	MRR	Number of Correct Response	MRR	Number of Correct Response
Definition	0.433	5/10	0.475	6/10	**0.570**	**7/10**	0.425	6/10
Why	0.377	9/17	0.345	9/17	**0.435**	**10/17**	0.240	6/17
How	0.222	2/3	0.261	**3/3**	0.317	**3/3**	0.111	1/3
Other	0.350	9/20	0.374	13/20	0.502	**14/20**	0.412	**14/20**
All	0.372	25/50	0.378	31/50	0.482	**34/50**	0.338	27/50

Table 3: Case of use of Similarity 2 (using question part only) in clustering examples Q&A pairs

	Proposed method ($\gamma = 0.7$)		Proposed method ($\gamma = 0.8$)		Proposed method ($\gamma = 0.9$)		Baseline ($\gamma = 0.5$)	
Type of Question	MRR	Number of Correct Response	MRR	Number of Correct Response	MRR	Number of Correct Response	MRR	Number of Correct Response
Definition	0.458	6/10	0.475	6/10	0.550	6/10	0.425	6/10
Why	0.325	8/17	0.355	8/17	0.422	9/17	0.240	6/17
How	0.511	**3/3**	0.178	2/3	0.4	2/3	0.111	1/3
Other	0.329	10/20	0.385	11/20	0.514	**14/20**	0.412	**14/20**
All	0.365	27/50	0.380	27/50	**0.483**	31/50	0.338	27/50

to the submitted question, move it from CL to CA.

3. For each cluster C in CL, calculate the generation probability of the submitted question on the model of question parts of $CA \cup \{C\}$, then find the cluster Cm that gives the maximum probability and move it from CL to CA.

4. Repeat the step 3 until the number of example questions in CA exceeds *target number*.

5.3 Extracting Answer Candidate of the Submitted Question Using the Probabilistic Language Model of Retrieved Example Answers

In this stage, we construct a langeuage model of example answers paired with example questions retrieved in Section 5.2. By Equation (5) in Section 5.1, according to the mixture probabilistic language model of part-of-speech tags and surface expressions, each sentence in answer candidates, which are retrieved by the same way as the baseline method in Section 3, are evaluated in terms of the appropriateness of writing style for the answers to the submitted question.

However, because of the nature of probability, the estimation of the appropriateness based on the probability unreasonably gives higher values to shorter sentences. Therefore, in order to resolve the problem, we normalized the Equation (5) as follows.

$$\bar{P}(E_1E_2...E_n) = \frac{1}{n}\log\{P(E_1E_2...E_n)\} \tag{6}$$

After the normalization, we calculate a score of the sentence S_i with Equation (7). We replace the last term in Equation (4) with Equation (6). Since the terms $\sum_{j=1}^{n} T(w_{ij})$ and $\bar{P}(E_1E_2...E_m)$ in Equation (7) correspond to Measure 1 and Measure 2, respectively, the parametar γ is used to determine the mixture ratio of Measure 1 and Measure 2.

$$\text{New Score}(S_i) = \frac{\left\{\sum_{j=1}^{n} T(w_{ij})\right\}^{\gamma}}{\log(1+|S_i|)} \cdot \left\{\bar{P}(E_1E_2...E_m)\right\}^{1-\gamma} \tag{7}$$

6 Experiments

We conducted some experiments to examine the effectiveness of the proposed method. In order to do it, we compared the system based on the proposed method with the system based on the baseline method described in Section 3. In the experiments, we especially investigated the dependence of the accuracy on the following two settings: 1) the value of parameter γ, which represents the mixture ratio

Table 4: Case of use of Similarity 3 (using answer part only) in clustering examples Q&A pairs

	Proposed method ($\gamma = 0.7$)		Proposed method ($\gamma = 0.8$)		Proposed method ($\gamma = 0.9$)		Baseline ($\gamma = 0.5$)	
Type of Question	MRR	Number of Correct Response	MRR	Number of Correct Response	MRR	Number of Correct Response	MRR	Number of Correct Response
Definition	0.458	6/10	0.483	6/10	0.500	6/10	0.425	6/10
Why	0.332	9/17	0.345	8/17	0.345	9/17	0.240	6/17
How	0.400	**3/3**	**0.611**	**3/3**	0.511	**3/3**	0.111	1/3
Other	0.527	13/20	**0.543**	14/20	0.502	**14/20**	0.412	**14/20**
All	0.439	31/50	0.464	31/50	0.437	32/50	0.338	27/50

of Measure 1 and Measure 2 in Equation (7) and 2) the similarity calculation methods in the clustering described in Section 5.2.

6.1 Experimental Settings

As the question set, we use the latter half of Japanese question set of NTCIR-6 QAC formal run test set (Fukumoto *et al.*, 2007).

As a Web search engine for information source of QA, we adopted Yahoo! Japan API[1]. With regard to Q&A examples, we used a corpus of 0.9 million Q&A pairs that comes from "Yahoo! Chiebukuro," which is a Q&A community site and the Japanese version of "Yahoo! answers." Let the parameter *target number* described in Section 5.2 be 500. The systems output five answers for each submitted question in the descending order of score. Judgment whether an answer candidate is correct or not is performed by one assessor. The assessor judged an output answer candidate correct, when the candidate includes correct answer for the question as its part. We use Mean Reciprocal Rank (MRR[2]) as the evaluation metrics. In addition to MRR, we also investigate the number of the questions for which the system can return, at least, one correct answer in the top five answer candidates (number of correct responses, hereafter).

6.2 Experimental Results

Experimental results are shown in the Table 2,3, and 4.

With regard to the baseline method, we employed 0.5 for the parameter γ, because it gives the best performance in terms of MRR. On the other hand, as for the proposed method, the results are shown for the three settings, γ =0.7,0.8,and 0.9, which give the better performance than other settings.

Although the proposed method and the baseline method do not perform any question classification, the results are shown on a type-by-type basis in order to investigate the effectiveness of the method for each typical question type described in Table 1.

6.3 Discussion

All of Table 2, 3, and 4 show that the proposed method outperforms the baseline method.

With regard to the number of correct responses, the proposed method gives more correct responses than the baseline method except for the case of use of Similarity 1 (using both question part and answer part) and $\gamma = 0.7$.

With regard to MRR, the proposed method gives better performance than the baseline method for not only the average of all questions but also the average of each type of question. One of the reasons for the good performance may be the fact that the propose method can appropriately filter out ungrammatical expressions in answer candidates, while the baseline method sometimes employ them as answer response. It means that the introduced probabilistic language model contribute to removing ungrammatical text from answer candidates. Another one of the reasons for the good performance may be the fact that the proposed method can reduce the number of example Q&A pairs which include unsuitable expressions for answers of the submitted question when the system retrieves example Q&A pairs. It means that more example answers suitable to the submitted question can be retrieved by introducing the clustering and the probabilistic estimation to the process of retrieving example questions, and as a result, by refining the language model of answers. The following shows an example for which the baseline method cannot give correct answer, but the proposed method can.

[1] http://developer.yahoo.co.jp/

[2] Reciprocal Rank (RR) is the inverse of the rank of the first correct answer candidate. MRR is the average of RRs over the question set.

Question (submitted)
What is required to effectuate the Kyoto Protocol? (Originally in Japanese)

Answer (Baseline)
After deposit of instrument of ratification of Kyoto protocol by the Russian goverment, a condition for ratification is satisfied, it is effectuated on February 16, 2005. (Originally in Japanese)

Answer (Proposed method)
In order to effectuate the Kyoto Protocol, the ratification by more than 50 signatory countries and contries whose carbon-dioxide emission is more than 55% of advanced industrial countries' are needed. (Originally in Japanese)

With regard to the methods of similarity calculation in clustering example Q&A pairs, Similarity 1 (using both question part and answer part) generally gives better performance than other similarity calculation methods in terms of both the number of correct response and MRR. The following reason may be supposed.

- The features from question parts of retrieved Q&A examples seem not to be suitable for clustering the Q&A examples because the writing styles of question parts are very similar to each other on account of the method for retrieving Q&A examples. In order to retrieve example questions similar to the submitted question, we use the 7-gram in each question part whose center word is an interrogative.
- Since answer parts have longer text than question parts in Q&A examples and are consequently described in various writing styles, it may be possible to find subgroups of answer parts according to the variations of writing styles.

For these reasons, the use of answer parts of Q&A examples is more efficient for clustering the examples. Although there is no significant difference between Similarity 1 and Similarity 3 (using answer part only) as shown in Table 2 and 4, the system with Similarity 1 (γ=0.9) stably outperforms the system with Similarity 3 in terms of the number of correct response. Moreover, MRR of the system with Similarity 1, 0.482, is almost the same as the best performance, 0.483, among all settings.

7 Conclusion

In this study, we poposed a method to introduce a probabilistic language model into non-factoid question answering in order to improve the accuracy ot the system proposed by Mori *et al.* (2008)

We introduced the model into two sub-processes which calculate similarity in terms of writing style. The first process collects example questions similar to an submitted question. The second one measures similarity between an answer candidate and example answers paired with the collected example questions. The experimental results showed that the system with the propose method outperforms the baseline system.

References

Fukumoto, J. 2007. Question Answering System for Non-factoid Type Questions and Automatic Evaluation based on BE Method. *Proceedings of the Sixth NTCIR Workshop Meeting*, Tokyo, Japan,441–447.

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

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

Heie, M.H., E.W.D. Whittaker and S. Furui. 2012. Question answering using statistical language modelling. *Computer Speech and Language 26*, 193–209.

Mizuno, J., T. Akiba, A. Fujii and K. Itou. 2009. Non-factoid Question Answering Experiments at NTCIR-6:Towards Answer Type Detection for Realworld Questions. *Proceedings of the Sixth NTCIR Workshop*, 487–492.

Mori, T., M. Sato and M. Ishioroshi. 2008. Answering any class of Japanese non-factoid question by using the Web and example Q&A pairs from a social Q&A website. *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 59–65.

Soricut, R., T. Akiba and E. Brill. 2006. Automatic Question Answering Using the Web: Beyond the Factoid. *Journal of Information Retrieval - Special Issue on Web Information Retrieval*, vol.9, 191–206.

Takahashi, A., A. Takatsu and J. Adachi. 2010. Language Model Combination for Community-based Q&A Retrieval. *Proceedings of the 2010 22nd IEEE International Conference on Tools with Artificial Intelligence*, 241–248.