Statistical Machine Translation based Passage Retrieval for Cross-Lingual Question Answering

Tomoyosi Akiba **Kei Shimizu**
Dept. of Information and Computer Sciences,
Toyohashi University of Technology
1-1 Hibarigaoka, Tenpaku-cho, Toyohashi-shi,
441-8580, JAPAN
akiba@cl.ics.tut.ac.jp

Atsushi Fujii
Graduate School of Library,
Information and Media Studies,
University of Tsukuba
1-2 Kasuga, Tsukuba, 305-8550, JAPAN
fujii@slis.tsukuba.ac.jp

Abstract

In this paper, we propose a novel approach for Cross-Lingual Question Answering (CLQA). In the proposed method, the statistical machine translation (SMT) is deeply incorporated into the question answering process, instead of using it as the pre-processing of the mono-lingual QA process as in the previous work. The proposed method can be considered as exploiting the SMT-based passage retrieval for CLQA task. We applied our method to the English-to-Japanese CLQA system and evaluated the performance by using NTCIR CLQA 1 and 2 test collections. The result showed that the proposed method outperformed the previous pre-translation approach.

1 Introduction

Open-domain Question Answering (QA) was first evaluated extensively at TREC-8 (Voorhees and Tice, 1999). The goal in the factoid QA task is to extract words or phrases as the answer to a question from an unorganized document collection, rather than the document lists obtained by traditional information retrieval (IR) systems. The cross-lingual QA task, which has been evaluated at CLEF (Magnini et al., 2003) and NTCIR (Sasaki et al., 2005), generalizes the factoid QA task by allowing the different languages pair between the question and the answer.

Basically, the CLQA system can be constructed simply by translating either the question sentence or the target documents into the language of the other side, and applying a mono-lingual QA system. For example, after the English question sentence is translated into Japanese, a Japanese mono-lingual QA system can be applied to extract the answer from the Japanese target documents. Depending on the translation techniques used for the pre-processing, the previous CLQA approach can be classified into the machine translation based approach (Shimizu et al., 2005; Mori and Kawagishi, 2005) and the dictionary based approach (Isozaki et al., 2005).

In this paper, we propose a novel approach for CLQA task. In the proposed method, the statistical machine translation (SMT) (Brown et al., 1993) is deeply incorporated into the question answering process, instead of using the SMT as the pre-processing before the mono-lingual QA process as in the previous work. Though the proposed method can be applied to any language pairs in principle, we focus on the English-to-Japanese (EJ) CLQA task, where a question sentence is given in English and its answer is extracted from a document collection in Japanese.

Recently, language modeling approach for information retrieval has been widely studied (Croft and Lafferty, 2003). Among them, statistical translation model has been applied for mono-lingual IR (Berger and Lafferty, 1999), cross-lingual IR (Xu et al., 2001), and mono-lingual QA (Murdock and Croft, 2004). Our method can be considered as that applying the translation model to cross-lingual QA.

In the rest of this paper, Section 2 summarizes the previous approach for CLQA. Section 3 describes our proposed method in detail. Section 4 describes the experimental evaluation conducted to see the performance of the proposed method by comparing it to some reference methods. Section 5 describes our conclusion and future works.

2 Previous CLQA Systems

Figure 1 shows the configuration of our previous English-to-Japanese cross-lingual QA system, which has almost the same configuration to the conventional CLQA systems. Firstly, the input English question is translated into the corresponding Japanese question by using a machine translation. Alternatively, the machine translation can be re-

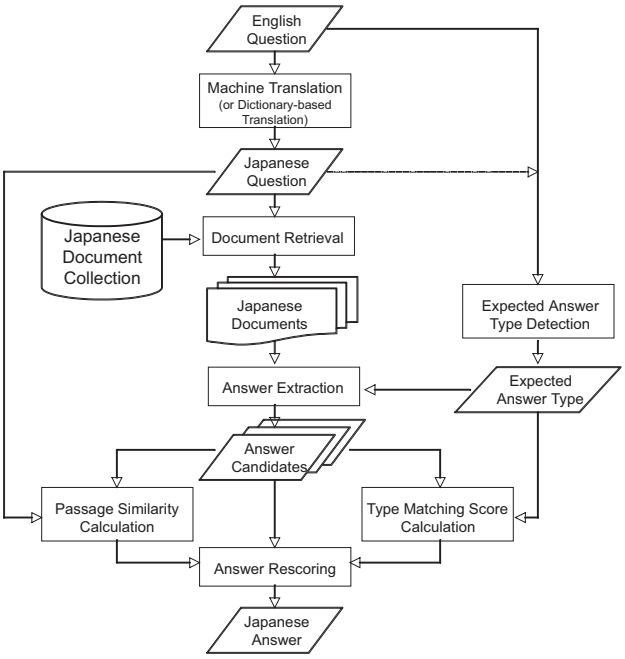

Figure 1: The configuration of the conventional CLQA system.

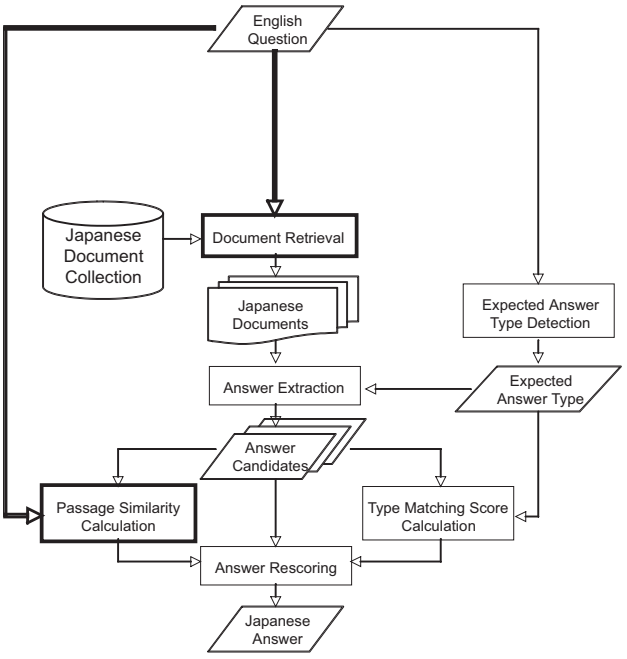

Figure 2: The configuration of the proposed CLQA system.

placed by the dictionary-based term-by-term translation. Then, either the English question or the translated Japanese question is analyzed to get the expected answer type.

After that, the mono-lingual QA process is invoked. The translated Japanese question is used as the query of the document retrieval to get the documents that include the query terms. From the retrieved documents, the answer candidates that match with the expected answer type are extracted with their location in the documents. Next, the extracted candidates are rescored by the two points of views; the passage similarity and the type matching. The passage similarity is calculated between the translated Japanese question and the Japanese passage that surrounds the answer candidate, while the type matching score is calculated as the likelihood that the candidate is matched with the expected answer type. Finally the reordered candidates are outputted as the answers of the given question.

3 Proposed CLQA System

On the other hand, Figure 2 shows the configuration of our proposed cross-lingual QA system. It does not use the machine translation (nor the dictionary-based translation) as the pre-processing of the input English question. The original English question is used directly in the QA process. In order to make this approach possible, the two subsystems, the document retrieval subsystem and the passage similarity calculation subsystem, which are pointed by the direct arrow from the English question and are emphasized by the thick frames in Figure 2, are *cross-lingualized* to accept the English question directly instead of the Japanese question, by means of incorporating the statistical machine translation (SMT) process deeply into them.

In the following two subsections, we will explain how these two subsystems can deal with the English question directly. The document retrieval subsystem is modified so that the Japanese documents are indexed by English terms. The word translation probability used in the SMT is used to index the Japanese document with the corresponding English terms without losing the consistency. The passage similarity calculation subsystem calculates the similarity between an English question and a Japanese passage in terms of the probability that the Japanese passage is translated into the English question.

3.1 Document Retrieval

Given an English question sentence, the document retrieval subsystem of our proposed CLQA system retrieves Japanese documents directly. In order to do so, each Japanese document in the target collection

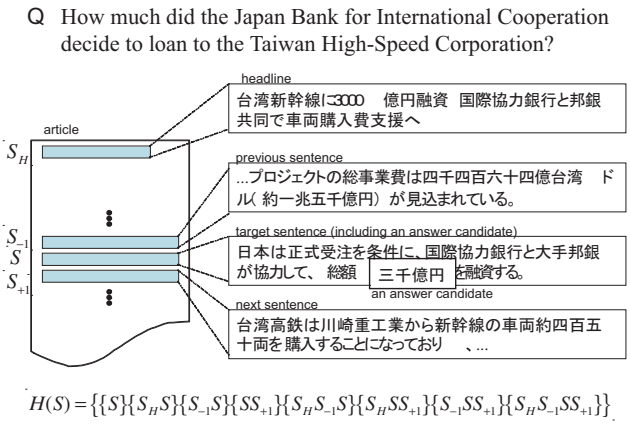

Figure 3: An examples of a question and the corresponding passage candidates.

has been indexed by English terms by using the word translation probability used in the SMT framework.

The expected term frequency $tf(e, D)$ of an English term e that would be used as an index to a Japanese document D can be estimated by the following equation.

$$tf(e, D) = \sum_{j \in D} t(e|j) tf(j, D) \quad (1)$$

where $tf(j, D)$ is the term frequency of a Japanese term j in D and $t(e|j)$ is the word translation probability that j is translated into e. The probability $t(e|j)$ is trained by using a large parallel corpus as the SMT framework. Because the expected term frequency $tf(e, D)$ is consistent with $tf(j, D)$ that is calculated from the statistics of D, the conventional vector space IR model based on the TF-IDF term weighting can be used for implementing our IR subsystem. We used *GETA* [1] as the IR engine in our CLQA system.

3.2 SMT based Passage Retrieval

In order to enable the direct passage retrieval, where the query and the passage are in different languages, the statistical machine translation is utilized to calculate the similarity between them. In order words, we calculate the similarity between them as the probability that the Japanese passage is translated into the English question.

The similarity $sim(Q, S|A)$ between a question Q and a sentence S including an answer candidate A is calculated by the following equation.

$$sim(Q, S|A) = \max_{D \in H(S)} P(Q|D - A) \quad (2)$$

[1] http://geta.ex.nii.ac.jp

where $P(Q|D - A)$ is the probability that a word sequence D except A is translated into a question sentence Q, and $H(S)$ is the set of the candidate passage (term sequences) that are related to a sentence S. The set consists of S and the power set of S_H, S_{-1}, and S_{+1}, where S_H is the headline of the article that S belongs, S_{-1} is the previous sentence of S, and S_{+1} is the next sentence of S (Figure 3).

In this paper, we use IBM model 1 (Brown et al., 1993) in order to get the probability $P(Q|D - A)$ as follows.

$$P(Q|D - A) = \frac{1}{(n+1)^m} \prod_{j=1}^{m} \sum_{i=1,\cdots,k-1,k+l+1,\cdots,n} t(q_j|d_i) \quad (3)$$

where $q_1 \cdots q_m$ is a English term sequence of a question Q, $d_1 \cdots d_n$ is a Japanese term sequence of a candidate passage D, $d_k \cdots d_{k+l}$ is a Japanese term sequence of an answer candidate A. Therefore, the Japanese term sequence $d_1, \cdots, d_{k-1}, d_{k+l+1}, \cdots, d_n$ (= D - A) is just D except A. We exclude the answer term sequence A from the calculation of the translation probability, because the English terms that corresponds to the answer should not be appeared in the question sentence as the nature of question answering.

4 Experimental Evaluation

The experimental evaluation was conducted to see the total performance of cross language question answering by using our proposed method.

4.1 Test collections

The NTCIR-5 CLQA1 test collection (Sasaki et al., 2005) and the NTCIR-6 CLQA2 test collection (Sasaki et al., 2007) for English-to-Japanese task were used for the evaluation. Each collection contains 200 factoid questions in English. The target documents for CLQA1 are two years newspaper articles from "YOMIURI SHINBUN" (2000-2001), while those for CLQA2 are two years articles from "MAINICHI SHINBUN" (1998-1999).

In the test collections, the answer candidates are judged with three categories; **Right**, **Unsupported**, and **Wrong**. The answer labeled **Right** is correct and supported by the document that it is from. The answer labeled **Unsupported** is correct but not supported by the document that it is from. The answer labeled **Wrong** is incorrect. We used two kind of golden set for our evaluation: the set including only

Right answers (referred as to **R**) and the set including **Right** and **Unsupported** answers (referred as to **R+U**).

Note that the evaluation results obtained from CLQA2 are more reliable than that from CLQA1, because we participated in CLQA2 formal run with our proposed method (and our reference method labeled **DICT**) and most of the answers by the system were manually checked for the pooling.

4.2 Translation Model

The translation model used for our method was trained from the following English-Japanese parallel corpus.

- 170,379 example sentence pairs from the Japanese-English and English-Japanese dictionaries.
- 171,186 sentence pairs from newspaper articles obtained by the automatic sentence alignment (Utiyama and hitoshi Isahara, 2003).

A part of the latter sentence pairs were obtained from the paired newspapers that are "YOMIURI SHINBUN" and its English translation "Daily Yomiuri". Because the target documents of CLQA1 are the articles from "YOMIURI SHINBUN" as described above, the corresponding sentence pairs, which are extracted from the articles from 2000 to 2001, were removed from the training corpus for CLQA1.

Before training the translation model, both English and Japanese sides of the sentence pairs in parallel corpus were normalized. For the sentences of Japanese side, the inflectional words were normalized to their basic forms by using a Japanese morphological analyzer. For the sentences of English side, the inflectional words were also normalized to their basic forms by using a Part-of-Speech tagger and all the words were lowercased. GIZA++ (Och and Ney, 2003) was used for training the IBM model 4 from the normalized parallel corpus. The vocabulary sizes were about 58K words for Japanese side and 74K words for English side. The trained Japanese-to-English word translation model $t(e|j)$ was used for our proposed document retrieval (Section 3.1) and passage similarity calculation (Section 3.2).

4.3 Compared methods

The proposed method was compared with the several reference methods. As the methods from previous works, three pre-translation methods were investigated.

The first two methods translate the question by using machine translation. One of them used a commercial off-the-shell machine translation software [2] (referred to as **RMT**). The other used the statistical machine translation that had been created by using the IBM model 4 obtained from the same parallel corpus and tools described in Section 4.2, the tri-gram language model constructed by using the target documents of CLQA1, and the existing SMT decoder (Germann, 2003) (referred to as **SMT**). The two methods, **RMT** and **SMT**, differ only in the translation methods, while their backend monolingual QA systems are common.

The third method translates the question by using translation dictionary (referred to as **DICT**). The cross-lingual IR system described in (Fujii and Ishikawa, 2001) was used for our "document retrieval" subsystem in Figure 2. The CLIR system enhances the basic translation dictionary, which has about 1,000,000 entries, with the compound words obtained by using the statistics of the target documents and with the borrowed words by using the transliteration method. Note that, as the other parts of the system than the document retrieval, including proposed SMT based passage retrieval, are all identical to the proposed method, this comparison is focused only on the difference in the document retrieval methods.

In order to investigate the performance if the ideal translation is made, the reference Japanese translations of the English questions included in the test collections were used as the input of the monolingual QA system (referred to as **JJ**).

As the variations of the proposed method, the following four methods were compared.

Proposed The same method as described in Section 3.

Proposed +r The document retrieval score is also used to rescore the answer candidates in "Rescoring" subsystem in Figure 2, in addition to the passage similarity score and the type matching score.

Proposed -p For the passage similarity calculation, the passage is always fixed only the central sentence S, i.e. the equation (2) is replaced by the following.

$$sim(Q, S|A) = P(Q|S - A) \qquad (4)$$

Proposed -p+r Combination of above two modifications.

[2] "IBM Japan, honyaku-no-oosama ver. 5"

Table 1: Comparison of the **JJ** results between the test collections.

test collection	**R**			**R+U**		
	Top1 Acc.	Top5 Acc.	MRR	Top1 Acc.	Top5 Acc.	MRR
CLQA1	0.140	0.300	0.196	0.260	0.535	0.354
CLQA2	0.245	0.410	0.307	0.270	0.530	0.366

Table 2: The performances of the proposed and reference CLQA systems with respect to CLQA1 test collection.

method	Top1 Acc.	Top5 Acc.	MRR
RMT	0.065	0.175	0.099
SMT	0.060	0.175	0.098
Dict	0.095	0.195	0.134
Proposed	0.090	0.225	0.146

Table 3: The performances among the proposed methods with respect to CLQA1 test collection.

method	Top1 Acc.	Top5 Acc.	MRR
Proposed	0.090	0.225	0.146
Proposed +r	0.105	0.285	0.173
Proposed -p	0.105	0.245	0.155
Proposed -p+r	0.120	0.280	0.178
JJ	0.260	0.535	0.354

4.4 Evaluation Metrics

Each system outputted five ranked answers $a_1 \cdots a_5$ for each question q. We investigated the performance of the systems in terms of three evaluation metrics that are obtained by averaging over all the questions: the accuracy of the top ranked answers (referred to as **Top 1 Acc.**), the accuracy of up-to fifth ranked answers (referred to as **Top 5 Acc.**), and the reciprocal rank (referred to as **MRR**) $RR(q)$ calculated by the following equation.

$$rr(a_i) = \begin{cases} 1/i & \text{if } a_i \text{ is a correct answer} \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

$$RR(q) = \max_{a_i} rr(a_i) \quad (6)$$

4.5 Results

Firstly, we compared the results obtained by using CLQA1 test collection with that obtained by using CLQA2. Table 1 shows the results for **JJ** system. By using the **R** judgment, the **JJ** results of CLQA1 was much worse than that of CLQA2, while the results were almost same by using the **R+U** judgment. Because the difference with respect to the difficulties between the two test collections seems small and the results from CLQA2 are more reliable, we concluded that the **R** judgment of CLQA1 was unreliable. Therefore, for CLQA1 test collection, we only investigated the result by using **R+U** judgment.

Secondly, we compared the proposed method (**Proposed**) with the previous methods (**RMT**, **SMT**, and **Dict**). Table 2 shows the results with respect to CLQA1 test collection. The two methods based on the machine translation (**RMT** and **SMT**) indicated almost same performance, while the performance of the proposed method was about 1.3 to 1.5 times better for CLQA1. Especially, because the same training data was used to build the translation models both in **SMT** and **Proposed**, it was shown that the method to build the SMT model in the QA process was better than that to use the same SMT model for pre-processing (pre-translating) the input sentence.

The **DICT** performed almost same as the **Proposed** for CLQA1, while **Proposed** was 1.7 to 1.9 times better than **DICT** for CLQA2 as shown in Table 4. Note again that this comparison was focused on the document retrieval subsystem, because the passage retrieval subsystems of these two methods were same.

Thirdly, the variations between the proposed methods were compared. Table 3 shows the results with respect to CLQA1 test collection. For CLQA1, both the additional use of the document retrieval score (**+r**) and the use of the fixed central sentence for passage similarity calculation (**-p**) improved the performance. However, for CLQA2, the document retrieval score (**+r**) did not contribute to improve the performance, as shown in Table 4.

Finally, seeing from the comparison between **JJ** and **Proposed**, it was shown that the performance of the proposed CLQA system was about half of that of the ideal CLQA system.

Table 4: The performances of the proposed and reference CLQA systems with respect to CLQA2 test collection.

methods	R			R+U		
	Top1 Acc.	Top5 Acc.	MRR	Top1 Acc.	Top5 Acc.	MRR
Dict	0.070	0.155	0.102	0.100	0.275	0.163
Proposed	0.130	0.200	0.155	0.165	0.295	0.210
Proposed +r	0.120	0.220	0.153	0.155	0.325	0.211
JJ	0.245	0.410	0.307	0.270	0.530	0.366

5 Conclusion

In this paper, a novel approach for CLQA was proposed. The proposed method did not translate the input question in source language into the target language as the preprocessing of QA process. Instead, the statistical machine translation was deeply incorporated into the two QA subsystems in order to deal with the question in source language directly in the QA process. Especially, SMT-based passage retrieval was explored.

For the passage similarity calculation in this paper, the simple IBM model 1 was used. In the future work, we will investigate if the more sophisticated translation model or that specialized for CLQA task can improve the performance further.

References

Adam Berger and John Lafferty. 1999. Information retrieval as statistical translation. In *Proceedings of the 22nd Annual Conference on Research and Development in Information Retrieval (ACM SIGIR)*, pages 222–229.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 18(4):263–311.

W. Bruce Croft and John Lafferty, editors. 2003. *Language Modeling for Information Retrieval*. Kluwer Academic Publishers.

Atsushi Fujii and Tetsuya Ishikawa. 2001. Japanese/english cross-language information retrieval: Exploration of query translation and transliteration. *Computers and the Humanities*, 35(4):389–420.

Ulrich Germann. 2003. Greedy decoding for statistical machine translatioin in slmost linear time. In *Proceedings of HLT-NAACL*.

Hideki Isozaki, Katsuhito Sudoh, and Hajime Tsukada. 2005. NTT's japanese-english cross-language question answering system. In *Proceedings of The Fifth NTCIR Workshop*, pages 186–193.

Bernardo Magnini, Alessandro Vallin, Christelle Ayache, Gregor Erbach, Anselmo Pe nas, Maarten de Rijke, Paulo Rocha, Kiril Simov, and Richard Sutcliffe. 2003. Overview of the CLEF 2004 multilingual question answering track. In *Multilingual Information Access for Text, Speech and Images*, pages 371–391.

Tatsunori Mori and Masami Kawagishi. 2005. A method of cross language question-answering based on machine translation and transliteration. In *Proceedings of The Fifth NTCIR Workshop*, pages 215–222.

Vanessa Murdock and W. Bruce Croft. 2004. Simple translation models for sentence retrieval in factoid question answering. In *Proceedings of the Workshop on Information Retrieval for Question Answering*.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Yutaka Sasaki, Hsin-Hsi Chen, Kuang hua Chen, and Chuan-Jie Lin. 2005. Overview of the NTCIR-5 cross-lingual question answering task (clqa1). In *Proceedings of The Fifth NTCIR Workshop*, pages 175–185.

Yutaka Sasaki, Chuan-Jie Lin, Kuang hua Chen, and Hsin-Hsi Chen. 2007. Overview of the NTCIR-6 cross-lingual question answering (clqa) task. In *Proceedings of The NTCIR-6 Workshop Meeting*.

Kei Shimizu, Tomoyosi Akiba, Atsushi Fujii, and Katunobu Itou. 2005. Bi-directional cross language question answering using a single monolingual QA system. In *Proceedings of The Fifth NTCIR Workshop*, pages 236–237.

Masao Utiyama and hitoshi Isahara. 2003. Reliable measures for aligning japanese-english news articles and sentences. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*, pages 72–79.

E. Voorhees and D. Tice. 1999. The TREC-8 question answering track evaluation. In *Proceedings of the 8th Text Retrieval Conference*, pages 83–106, Gaithersburg, Maryland.

Jinxi Xu, Ralph Weischedel, and Chanh Nguyen. 2001. Evaluating a probabilistic model for cross-lingual information retrieval. In *Proceedings of the 24th Annual Conference on Research and Development in Information Retrieval (ACM SIGIR)*, pages 105–110.