Statistical Machine Translation based Passage Retrieval for Cross-Lingual Question Answering — Experiments at NTCIR-6

Tomoyosi Akiba *Kei Shimizu*
Department 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

Katunobu Itou
Graduate School of Computer and Information Sciences, Hosei University
3-7-2 Kajino-cho, Koganei-shi, Tokyo, 387-6028, JAPAN

Abstract

In this paper, we propose a novel approach for Cross-Lingual Question Answering (CLQA), where the statistical machine translation (SMT) is utilized. In the proposed method, the 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. Our experimental results targeting the English-to-Japanese CLQA using the NTCIR CLQA 1 and 2 test collections 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 [16]. 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 evaluated in the NTCIR CLQA task series generalizes the factoid QA task by allowing the different languages pair between the question and the answer.

Basically, the CLQA system can be realized by making the languages uniform between the given question and the target document collection by applying an automatic translation technique. For example, after the question sentence is translated to the same language as the target document collection, a mono-lingual QA system can be applied to get the answer of the question. Depending on the translation techniques used for the unification, the previous CLQA approach can be classified into the machine translation based approach [14, 9] and the dictionary based approach [8].

In this paper, we propose a novel approach for CLQA task, where the statistical machine translation [3] is utilized. In the proposed method, the statistical machine translation (SMT) is deeply incorporated into the question answering process, instead of using the SMT as the pre processing of the mono-lingual QA process as in the previous work. Though the proposed method can be applied to any language pairs in principle, we focused on the English-to-Japanese (EJ) CLQA task, where a question sentence is given in English and its answer should be extracted from a document collection in Japanese.

In Section 2, we relate our approach to the previous works. 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 conclution and future works.

2 Related Works

Recently, language modeling approach for information retrieval has been widely studied [4]. Among them, statistical translation model has been applied for mono-lingual IR [2], cross-lingual IR [17], and mono-lingual QA [10]. Our method can be seen as in the

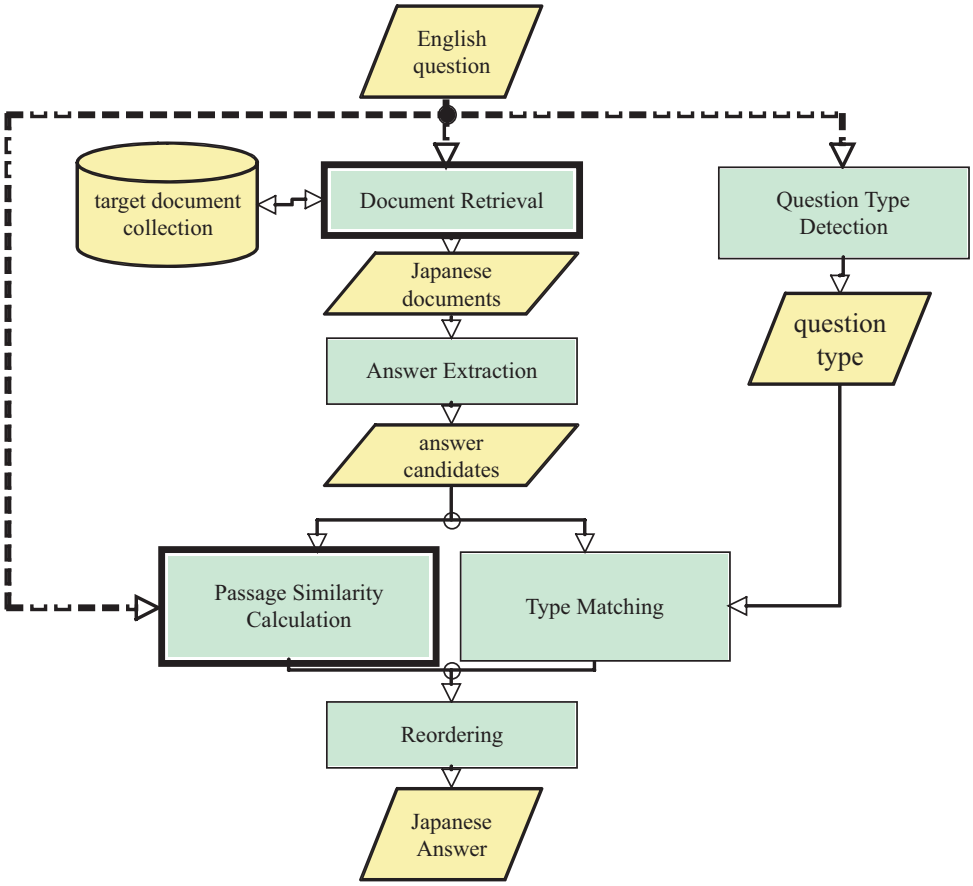


Figure 1. The configuration of the proposed CLQA system.

line of these works and applying the translation model to cross-lingual QA.

3 Proposed CLQA system

Figure 1 shows the configuration of our cross-lingual QA system. The system has the same structure as our mono-lingual Japanese QA system [1] excepting that the two subsystems, the document retrieval subsystem and the passage similarity calculation subsystem, which are emphasized by the thick frames in Figure 1. They are modified to deal with an input English question sentence directly instead of a Japanese question sentence as in the original mono-lingual QA system.

The document retrieval subsystem enables to retrieve Japanese documents directly from an input English question sentence by means of indexing the Japanese document collection with English terms.

The passage similarity calculation subsystem calculates the similarity between an input English question sentence and Japanese passage around an answer candidate by means of the probability of translating the context sentences into the question.

3.1 Document Retrieval

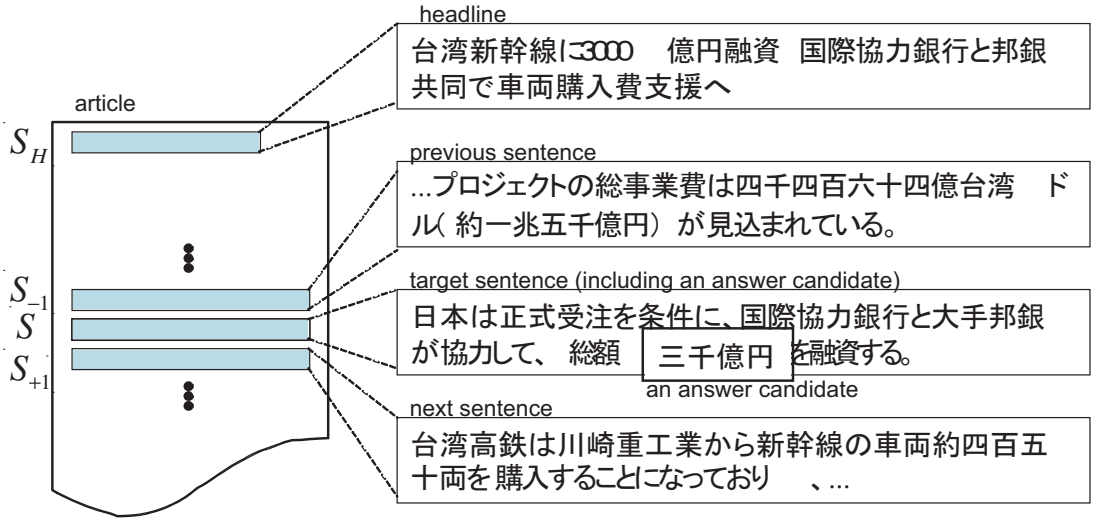
Given an English question sentence, the document retrieval subsystem of our proposed CLQA system retrieves Japanese documents directly. In order to do so, each Japanese document in the target collection has been indexed by English terms by using the translation model of statistical machine translation.

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

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

where $t(j, D)$ is the term frequency of a Japanese terms j in D and $t(e|j)$ is the translation probability that j is translated to e. The translation model $t(e|j)$ can be obtained from large parallel corpus as in the framework of statistical machine translation. 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* [7] as the

Q How much did the Japan Bank for International Cooperation decide to loan to the Taiwan High-Speed Corporation?



$$H(S)=\{\{S\}\{S_HS\}\{S_{-1}S\}\{SS_{+1}\}\{S_HS_{-1}S\}\{S_HSS_{+1}\}\{S_{-1}SS_{+1}\}\{S_HS_{-1}SS_{+1}\}\}$$

Figure 2. An Examples of a question and the corresponding passage candidates.

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 language each other, 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 to 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.

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

where $P(Q|D)$ is the probability that a word sequence D is translated to 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 following sentence of S (Figure 2).

In this paper, we used IBM model 1 [3] in order to model the probability $P(Q|D)$ as follows.

$$P(Q|D-A) = \frac{1}{(n+1)^m} \prod_{j=1}^{m} \sum_{i=1,2,\cdots,k-1k+l+1,\cdots,n} t(q_j|d_i) \qquad (3)$$

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 excepting 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 [12] and the NTCIR-6 CLQA2 test collection [13] for English-to-Japanese task were used for the evaluation. Each collection contains 200 factoid questions in English. The target document for CLQA1 is two years newspaper articles of "YOMIURI SHINBUN" (2000-2001),

while that for CLQA2 is two years newspaper articles of "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**).

4.2 Translation Model

The translation model used for our method was obtained by training 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 [15].

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 the alphabets of all words were transformed to small letters. GIZA++ [11] 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 obtained Japanese-to-English word translation probabilities $P(e|j)$ were 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 three differ only in the translation methods, while their backend mono-lingual QA system is common.

The first two methods translate the question by using machine translation. One of them used a commercial off-the-shell machine translation software (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 [6] (referred to as **SMT**).

The third method translates the question by using translation dictionary (referred to as **DICT**). The cross-lingual IR system described in [5] was used for our "document retrieval" subsystem in Figure 1. 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 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 mono-lingual 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 Rescoring the answers in "Rescoring" subsystem in Figure 1 by not only the passage similarity and the type matching but also the document retrieval 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.

4.4 Results

Each system outputted five ranked answers $a_1 \cdots a_5$ according to the score for each question q. We investigated the performance of the systems by means of three evaluation metrics averaged over the questions: the accuracy of the top ranked answers

Table 1. The performances of the proposed and reference CLQA systems with respect to CLQA1 test collection.

methods	R			R+U		
	Top1 Acc.	Top5 Acc.	MRR	Top1 Acc.	Top5 Acc.	MRR
JJ	0.140	0.300	0.196	0.260	0.535	0.354
RMT	0.020	0.075	0.039	0.065	0.175	0.099
SMT	0.015	0.070	0.034	0.060	0.175	0.098
Dict	0.055	0.100	0.070	0.095	0.195	0.134
Proposed	0.045	0.125	0.074	0.090	0.225	0.146
Proposed +r	0.050	0.140	0.083	0.105	0.285	0.173
Proposed -p	0.040	0.120	0.069	0.105	0.245	0.155
Proposed -p+r	0.055	0.155	0.091	0.120	0.280	0.178

Table 2. 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
JJ (TTH-J-J-u-01)	0.245	0.410	0.307	0.270	0.530	0.366
Dict (TTH-E-J-u-03)	0.070	0.155	0.102	0.100	0.275	0.163
Proposed (TTH-E-J-u-01)	0.130	0.200	0.155	0.165	0.295	0.210
Proposed +r (TTH-E-J-u-02)	0.120	0.220	0.153	0.155	0.325	0.211

(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**) $R(q)$ calculated by the following equation.

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

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

Table 1 and 2 shows the results for CLQA1 and CLQA2, respectively.

Firstly, we compared the results obtained by using CLQA1 test collection with that obtained by using CLQA2. 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, 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. 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 preprocessing (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. 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. 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.

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.

5 Conclusion

In this paper, the statistical machine translation based passage retrieval was proposed. The CLQA system that the proposed method was built in outperformed the previous post-translation methods, which used the machine translation or the translation dictionary for converting the input question in source language to the target language as the preprocessing of QA process.

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

References

[1] T. Akiba, A. Fujii, and K. Itou. Question answering using "common sense" and utility maximization principle. In *Proceedings of The Fourth NTCIR Workshop*, 2004.

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

[3] P. F. Brown, S. A. D. Pietra, V. J. D. Pietra, and R. L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 18(4):263–311, 1993.

[4] W. B. Croft and J. Lafferty, editors. *Language Modeling for Information Retrieval*. Kluwer Academic Publishers, 2003.

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

[6] U. Germann, M. Jahr, K. Knight, D. Marcu, and K. Yamada. *ISI Rewrite Decoder User's Manual*. `http://www.isi.edu/natural-language/software/decoder/manual.html`.

[7] GETA. Generic engine for transposable association (GETA). `http://geta.ex.nii.ac.jp`.

[8] H. Isozaki, K. Sudoh, and H. Tsukada. NTT's japanese-english cross-language question answering system. In *Proceedings of The Fifth NTCIR Workshop*, pages 186–193, 2005.

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

[10] V. Murdock and W. B. Croft. Simple translation models for sentence retrieval in factoid question answering. In *Proceedings of the Workshop on Information Retrieval for Question Answering*, 2004.

[11] F. J. Och. *GIZA++: Training of statistical translation model*. `http://www-i6.informatik.rwth-aachen.de/Colleagues/och/software/GIZA++.html`.

[12] Y. Sasaki, H.-H. Chen, K. hua Chen, and C.-J. Lin. Overview of the NTCIR-5 cross-lingual question answering task (clqa1). In *Proceedings of The Fifth NTCIR Workshop*, pages 175–185, 2005.

[13] Y. Sasaki, C.-J. Lin, K. hua Chen, and H.-H. Chen. Overview of the NTCIR-6 cross-lingual question answering (clqa) task. In *Proceedings of The NTCIR-6 Workshop Meeting*, 2007.

[14] K. Shimizu, T. Akiba, A. Fujii, and K. Itou. Bidirectional cross language question answering using a single monolingual QA system. In *Proceedings of The Fifth NTCIR Workshop*, pages 236–237, 2005.

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

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

[17] J. Xu and W. B. Croft. 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, 2001.