A Method of Cross Language Question-Answering Based on Machine Translation and Transliteration
— Yokohama National University at NTCIR-5 CLQA1 —

Tatsunori MORI and Masami KAWAGISHI
Graduate School of Environment and Information Sciences, Yokohama National University
79-7 Tokiwadai, Hodogaya, Yokohama 240-8501, Japan
{mori,kawagisi}@forest.eis.ynu.ac.jp

Abstract

We propose methods of English-Japanese and Japanese-English cross language question-answering (E-J/J-E CLQA) that use machine translation (MT), transliteration, and an existing Japanese QA system that deeply depends on Japanese. With regard to the E-J task, in order to compensate the insufficiencies in the bilingual dictionary of an MT system, we utilize the documents on the Web to translate proper nouns. We also introduce a pattern-matching-based question type detection in the source language in order to cope with translation errors. With regard to the J-E task, we adopt a standard CLIR technique, namely, the bilingual-dictionary-based keyword translation in order to retrieve English documents relevant to Japanese keywords. In the translation, we utilize the information about the classification of translation in the bilingual dictionary as well as the probability of word occurrence obtained from a corpus. We also introduce the answer mapping method, which finds an English expression in the English documents that corresponds to each Japanese answer candidate.

The experimental result shows that our proposed method improves the accuracy of E-J/J-E CLQA.

Keywords: *E-J/J-E CLQA, machine translation, transliteration by the Web, answer mapping.*

1 Introduction

In recent years, the *question answering* (QA) has gained attention as a way of information access to a large amount of text. QA is the technology that extracts answers for user's natural-language question from a knowledge resource, i.e., a large amount of text.

Since the knowledge resource may be a collection of documents from all over the world, the *cross-language* version of QA (CLQA) becomes one of important topics in the research area. CLQA is a task to answer to a given question written in a language by using a collection of documents written in other languages.

In this paper, we report the evaluation results of our CLQA system at NTCIR-5 CLQA1. We participated in the English-Japanese (E-J) task and the Japanese-English (J-E) task with two systems.

2 Related work

The Cross Language Evaluation Forum (CLEF)[3] introduced a new track termed *QA@CLEF* to test CLQA systems in 2003. At CLEF, there are several proposals about CLQA among European languages. In general, the methodologies of the proposals consist of the following steps: 1) translate a given question (or keywords in the question) to a target language using an MT system or bilingual dictionaries, 2) perform passage (or document) retrieval and linguistically analyze the retrieved passages, 3) assign a score to each answer candidate according to the degree of matching between the question and the retrieved passage. Although an NE recognizer is usually adopted as the linguistical analysis of retrieved passages, some systems extract semantic representation of sentences by a more deeper semantic analysis[1]. The documents on the Web is also exploited in some systems. For example, some of them validate the extracted answer candidate by using the Web[6]. The other system employs answer candidates obtained from the Web as well as the candidates in the document collection to be considered[2]. Roughly speaking, the processes following the query analysis may be regarded as a process of mono-lingual QA in the target language. Therefore, we have to introduce another (mono-lingual) QA system when we want to treat text in a new target language.

In order to address the problem, Sasaki[9] proposed a method of QA using a machine learning as a basis of CLQA. It does not need any named entity (NE) recognizer, which usually heavily depends on a certain language. However, the method requires a large amount of pairs of questions and their answers with context in cross language style when it is applied to CLQA. It also needs a cross-language IR system to reduce the number of documents to be processed by the method.

Another way to address the problem is the method to translate the collection of documents into the source language. However, it is pointed out that the method is not plausible[2]. The reason is that we have to have N different (translated) document collections for N languages. Moreover, the method is not applicable when the whole collection of documents is not in the control of CLQA systems like the Web.

To the contrary, our proposal is an attempt to construct CLQA systems based on only one mono-lingual QA engine. We employ the Japanese QA system we developed[7]. It has a special feature that it needs no preprocessing on documents and can handle unseen documents from external search engines. If we can translate a few dozen of passages into the source language with an MT system on the fly, it would be possible to perform CLQA with the Japanese QA system even if the task is J-E.

We also utilize the documents on the Web to translate proper nouns. With regard to transliteration by using the Web, Tsuji et al.[10] proposed a method of transliterating persons' names by using a Web search engine. It first generates candidates of transliteration by using the patterns learned from bilingual dictionaries, then sort them according to the number of hits on

the search engine. On the other hand, our method first searches for the documents that contain a target expression with a Web search engine, then finds the the expression and its translation from the snipped returned by the search engine. The method can find not only transliterations but also other type of translations.

3 Our approach

When we construct a CLQA system, we have to consider the following points.

When and how does the system translate text? Since a question and a collection of documents are written in different languages in CLQA, the system has to perform the translation of text to make them comparable to each other.

What language does the QA engine depend on? QA systems are usually language-dependent. Therefore, many of previous researches incorporate separate QA systems on a language-by-language basis.

The development of QA systems is a very laborious work. On the other hand, with regard to the matter of translation, many off-the-shelf MT products are available in the market. However, in general, the quality of output of MT is not enough for the basis of CLQA. Especially, some proper nouns, which convey very important information, are not translated properly because of the lack of vocabulary. Based on the above consideration, we are developing E-J and J-E CLQA systems according to the following policy.

- We adopt only one language-dependent QA system. In this paper, we use an existing Japanese QA system that deeply depends on Japanese.
- We change the place of MT stage(s) in the CLQA process dependently on the tasks.
- We utilize the documents on the Web to translate proper nouns that are not translated by MT.

In the rest of this section, we describe the Japanese QA system we adopted. Then, we will propose the E-J and J-E CLQA systems in the following sections.

As shown in Figure 1, the Japanese QA system consists of five modules, i.e., the question analyzer, the search engine, the passage extractor, the sentential matcher and the answer generator. The question

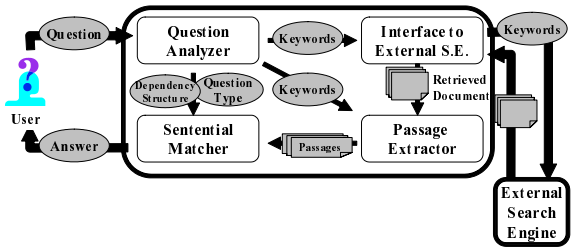

Figure 1. Overview of the Japanese QA system

analyzer receives a question from a user and extracts a list of keywords, the question type, and the dependency structure. Here, we define the term *keywords* as content words in a given question.

Since the information related to a question is usually contained in a very small part of the document, the passage extractor segments each document, which is retrieved by an external search engine, into small passages and selects suitable passages that are related to keywords. In our experiment, we defined one passage as a sequence of three sentences, similar to Murata et al.[8].

The sentential matcher receives a set of sentences in retrieved passages. The module treats each morpheme as an answer candidate and assigns it a matching score. The matching score represents the fitness of each answer candidate for the answer. We adopt a composite matching score shown in Equation (1), which is a linear combination of the following sub-scores for an answer candidate AC in the i-th retrieved sentence L_i with respect to a question sentence L_q: 1) the matching score in terms of 2-grams, $Sb(AC, L_i, L_q)$, 2) the matching score in terms of keywords, $Sk(AC, L_i, L_q)$, 3) the matching score in terms of dependency relations between an answer candidate and keywords, $Sd(AC, L_i, L_q)$, and 4) the matching score in terms of the question type, $St(AC, L_i, L_q)$. In the calculation of $St(AC, L_i, L_q)$, we employ an NE recognizer that spots NEs in eight types defined in the IREX-NE task[4].

$$\begin{aligned} & S(AC, L_i, L_q) \\ = \; & Sb(AC, L_i, L_q) + Sk(AC, L_i, L_q) \\ & + Sd(AC, L_i, L_q) + St(AC, L_i, L_q) \end{aligned} \quad (1)$$

4 Proposed method: CLQA for the E-J task

In the E-J task, we may use the passage extractor, the sentential matcher, and the answer generator for Japanese without any modifications, because in the task the target language is Japanese and the collection of documents is written in Japanese. On the other hand, the question analyzer should be revised so as to analyze English questions. There are at least two choices of the ways of revision as follows.

- Translate the (English) questions into Japanese by an MT system, then input the translated (Japanese) questions to the original question analyzer for Japanese.
- Newly construct a question analyzer for English questions that determines question types and extracts keywords. Translate the extracted keywords by using bilingual dictionaries.

In the former method, the system may be easily constructed by using an off-the-shelf MT system, but the system may encounter untranslated words and errors in translation, then consequently fail to extract proper keywords and detect question types. On the other hand, in the latter method, the system may exactly detect question types, but we have to explicitly introduce some context analysis in order to disambiguate the translation of words. Moreover, some important features in the Japanese QA system do not function, and it may cause some loss of accuracy. For example, dependency structures in Japanese questions are not available in the calculation of matching score.

Therefore, we proposes a method of question analysis that is a combination of the two method described above. The method is basically the first choice. However, with regard to the problem of untranslated words, we introduce a method to translate proper nouns by using the documents on the Web. With regard to the problem of detection of question types, we introduce a pattern-matching-based question type detector for the source language in order to cope with translation errors. Figure 2 shows the overview of the E-J CLQA system we proposed.

When a user submits a English question to the system, the system first translate it into Japanese with an MT system. In parallel with the translation, it detects the question type from the original English question.

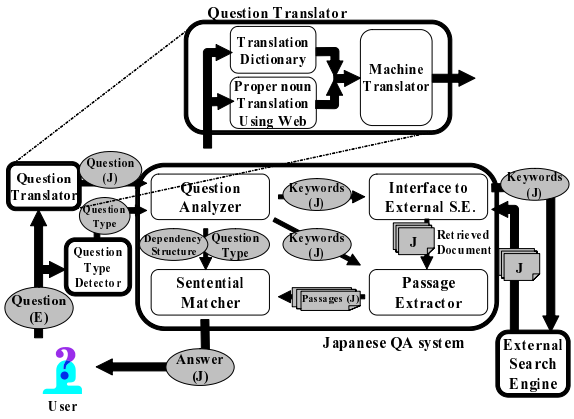

Figure 2. Overview of the proposed E-J CLQA system

The translated question is inputted to the question analyzer along with the detected question type. The rest of process is the exactly same as the Japanese QA.

4.1 Translating proper nouns using the Web

The following example is a question in the sample data of NTCIR-5 CLQA1 EJ task.

(1) What is the title of the book written by Roger Dingman about the Awa Maru case?

When we translate the question into Japanese with an MT system[1], the proper nouns, "Dingman" and "Awa Maru", remain untranslated in the translated question. The first word should be translated into a Katakana word[2] that is a transliteration of "Dingman." On the other hand, the words "Awa Maru" is the transliteration of pronunciation of a Japanese name that is written in Kanji characters. Therefore, we have to translate them into a string of Kanji characters. It is a difficult problem beyond the transliteration because a phoneme may corresponds to many Kanji characters. However, if we abandon these words, the QA system cannot find any correct answers.

In order to address the problem, we propose a method of extracting a pair of bilingual expression of one thing from Web documents. First, from an English question, we extract strings that may be proper nouns by the following procedure. The procedure detects originally Japanese proper nouns and other English proper nouns.

1. Find proper nouns that are capitalized or are not in the translation dictionary. Try to convert each proper noun into a string of Hiragana characters according to the Romanization rules of Japanese. If it is successful and the length of the converted string is more than one character, then the proper noun is judged an originally Japanese word.
2. If there is a sequence of capitalized words that are not of the above case, the sequence is detected as an English proper noun.

For example, the strings "Awa Maru" and "Roger Dingman" in Question (1) are detected as Japanese and English proper nouns respectively.

[1]We used Hon'yaku-no Ousama Ver.5[5].

[2]Katakana characters and Hiragana characters are Japanese phonograms. Loan words are usually transliterated into Katakana words. On the other hand, many Japanese words are written in Kanji characters. Since Kanji characters are Japanese ideograms, they may have more than one pronunciations.

4.1.1 Translation of strings detected as originally Japanese proper nouns

A string detected as an originally Japanese proper noun is translated into a Japanese string as follows.

1. Perform a Web search by using the detected string as a query and obtain a set of snippets from the search engine.
2. Remove alphabet characters and special characters from the set of snippets.
3. Apply a Japanese morphological analyzer to the set of snippets. Romanize each morpheme according to its pronunciation estimated by the morphological analyzer.
4. Extract each sequence of morphemes whose Romanized string is identical to the detected string. The set of the sequences may be Japanese translations that correspond to the detected string.

It should be noted that the method can collect translation candidates that may have the same pronunciation but have different expressions in Kanji characters.

4.1.2 Translation of strings detected as English proper nouns

A string detected as an English proper noun is also translated into a Japanese word by using snippets returned by a Web search engine. Since the snippets are usually sequences of sentence fragments and, in this case, contain both English expressions and Japanese expressions, it is difficult to precisely analyze the sentence structure around the target English strings. Therefore, we use a simple patter-match based method to translate the English string as follows. Here, we suppose that an English proper noun and its Japanese translation appear closely and in typical patterns if a document contains both of them.

1. Perform a Web search by using the detected string as a query.
2. Remove special characters from the set of snippets.
3. Apply the following patterns to the snippets and extract candidates of translation:
   - (*D*)?*SW*D*TW*D
   - *DTW*D*SW*D

where the symbol *D*, *SW*, and *TW* represent one of delimiting characters, the English proper noun, the target string to be extracted. The notation "(. . .)?" means that the pattern in the parentheses may be omitted.

4.1.3 Resolution of ambiguity in translation

When there are multiple candidates of translation for a proper noun, we generate multiple translations of the question according to the candidates. For example, Question (2a) in English is translated by an MT system into Question (2b) in Japanese where the word "*Sekigahara*" remains untranslated.

(2) a. When did the Battle of Sekigahara begin?

b. Itsu *Sekigahara*-No Tatakai-wa
WHEN Sekigahara-REL battle-TOP
Hajimari-mashi-ta-ka?
begin-POL-PAST-INTERROG
いつＳｅｋｉｇａｈａｒａの戦いは，始まりましたか？

The procedure described before detects the word "*Sekigahara*" as an originally Japanese proper noun, and extracts the translation candidates, i.e., "関ケ原", "関ヶ原", "せきがはら", and "関が原," from the Web. In this case, we generate the following four questions:

(3) a. いつ関ケ原の戦いは，始まりましたか？

b. いつ関ヶ原の戦いは，始まりましたか？

c. いつせきがはらの戦いは，始まりましたか？

d. いつ関が原の戦いは，始まりましたか？

All of these questions are submitted to the Japanese QA system, and the results, i.e., the lists of answer candidates are merged according to scores of answer candidates.

4.2 Detection of question types in the source language

We introduce a set of pattern-matching-based rules to detect question types in the source language, i.e. English. The rules are manually developed by referring to the NTCIR-5 CLAQ1 EJ Sample Data. The accuracy of the question type detection is 86.7% for the developing data (i.e. closed test).

5 Proposed method: CLQA for the J-E task

The J-E task is totally different from the E-J task in terms of using the Japanese QA system. We may use the question analyzer of the QA system. However, the processes following the information retrieval need the help of an MT system because retrieved documents are written in English.

There are at least the following approaches to perform the J-E task with the Japanese QA system.

1. *Off-line document translation approach:*
   Translate the entire collection of (English) documents into Japanese in advance.
2. *Online partial translation approach:*
   First perform the CLIR to retrieve English documents relevant to a Japanese question. Then,
   (a) Translate the retrieved documents into Japanese, and perform the Japanese QA. The processes after the passage retrieval are the same as the Japanese QA.
   (b) Perform the passage retrieval also on English documents to retrieve English passages. Translate the passages into Japanese. The processes after the sentence matching are the same as the Japanese QA.

Approach 1 has the advantage that we may use the Japanese QA system with no modification. The translation of questions is also unnecessary. However, it is very time-consuming and costly because we have to translate the entire collection.

On the other hand, Approach 2 does not have the disadvantage, but it needs the question translation to perform the CLIR as well as the online translation of retrieved documents/passages. Approach 2a takes a relatively long time to translate all of retrieved documents on-the-fly, although we can use the Japanese QA system for the processes after the passage retrieval. With regard to Approach 2b, we cannot perform any processes specific to Japanese in the passage retrieval, but the computational cost of translation is much less than Approach 2a. It should be noted that the above approaches generate *Japanese answers* for a Japanese question because the knowledge source is translated into Japanese. However, in the J-E task, the system has to present the answers as *English expressions* in the original English documents. According to our preliminary experiment, the result of MT of a Japanese answer is sometimes different from the English answer in the original documents. Therefore, we have to map the Japanese answer to the original English answer by using the MT.

In this paper, we propose a J-E CLQA system based on Approach 2b. Figure 3 shows the overview of the system.

In this figure, the question analyzer is the same as the Japanese QA system. The analyzer extracts keywords from a Japanese question. The keywords are translated into English keywords based on a bilingual dictionary.

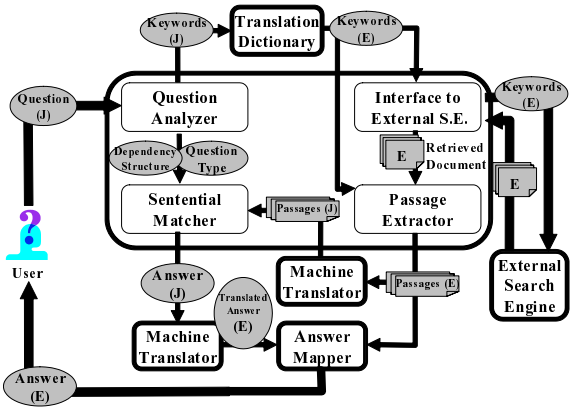

Figure 3. Overview of the proposed J-E CLQA system

The keywords translated into English are inputted to the search engine and the passage extractor, while the original Japanese keywords are passed to the sentential matcher. English passages related to the English keywords are retrieved from English documents. Then the passages are translated into Japanese. The sentential matcher finds answer candidates based on the result of analysis of a Japanese question and the original Japanese keywords along with the translated Japanese passages.

At the final stage, the answer mapping module finds an English expression in the English documents that corresponds to each answer candidate produced by the sentential matcher.

5.1 Translation of keywords

In the J-E task, the system does not need the translation of questions because questions are written in Japanese and the base QA system accepts Japanese questions. Only English-translated keywords are necessary in order to retrieve English documents/passages. Therefore, we take a keyword translation approach using a bilingual dictionary on a word-by-word basis. In general, the translation with a bilingual dictionary is ambiguous because one word may have more than one translation candidates. In order to address to the problem, in this paper, we assign a weight to translation candidates. The weight represents the appropriateness of a translation candidate.

There are many probabilistic ways to assign a weight to disambiguate translation. However, they usually need a bilingual corpus to obtain the information of translation probability. Therefore, we take a simpler approach based on a bilingual dictionary. we employ the EDR J-E dictionary as a bilingual dictionary. It has the information about the classification of translation as follows. Since the classification can be interpreted as the plausibility of a translation candidate, we assign a weight to each translation candidate according to the class as follows.

Equivalence The weight is 1.0. If the English translation is the same as the English head-concept of the Japanese word's entry, we add the extra weight 0.2 to the weight.

Paraphrase The weight is 0.9.

Word-for-word translation The weight is 0.7.

Romanization The English translation is the reading of the Japanese word spelled out in Roman letters. The weight is 0.65.

Explanation An explanation in English of the Japanese word is given. The weight is 0.3.

According to the weight we select the N-best translation. We term this method "*EDR JE only*."

Since the probability of word occurrence is not taken into account, the method sometimes assigns unintuitive weights. Therefore, we introduce the other method that combines the information of probability with the method "EDR JE only." We term this method "*EDR JE+Prob.*" The weight $W_2(tc_n)$ for the translation candidate tc_n is given as follows.

$$W_2(tc_n) = W_1(tc_n) \cdot P(tc_n)$$

$$P(tc_n) = \begin{cases} \frac{fr(tc_n)}{\sum_i fr(tc_i)+\beta} & (\text{if } fr(tc_n) \neq 0) \\ \frac{\beta}{\sum_i fr(tc_i)+\beta} \frac{1}{\sum_i holds(fr(tc_i)=0)} & (\text{if } fr(tc_n) = 0) \\ \gamma & (\text{if } tc_n \in \text{Stopwords}) \end{cases}$$

where $W_1(tc_n)$ is the weight for tc_n calculated by the method "EDR JE only", $fr(tc_n)$ is the frequency of tc_n in the collection of document. The function $holds()$ returns 1 when the argument is true, otherwise 0. β and γ are the parameters.

5.2 Answer mapping

The answer mapping module finds an English expression in the English documents that corresponds to each (Japanese) answer candidate produced by the sentential matcher as follows.

1. Translate the (Japanese) answer candidate into English by MT. The English-translated answer candidate usually consists of several English words. We term these words "*cue words*," and let n_{cw} be the number of cue words.
2. Mark all words in the original retrieved English documents/passages that satisfy one of the following conditions.
   (a) The length of the word is less than four characters, and the word is identical to one of cue words.
   (b) The length of the word is more than three characters, and one of the prefixes of the word is identical to one of cue words.
3. Chunk adjacent marked words as one marked phrase, then find all sequences of words that satisfy one of the following conditions. We term them *the strict mapping* and *the relaxed mapping*, respectively.
   (a) The sequence is a marked phrase and the number of words of it is equal to n_{cw}. (*strict mapping*)
   (b) The sequence begins with a marked phrase and ends with another marked phase, and the difference between n_{cw} and the number of words in the sequence is less than two.

6 Experimental Result 1: E-J Task

We use the sample test set (300 questions) and the formal run test set (200 questions) of NTCIR-5 CLQA1 E-J task as the data for developing the systems and the evaluation, respectively. With respect to our J-J reference run, the question set (200 questions) of the formal run of NTCIR-5 CLQA1 J-E task. As a machine translation system and a bilingual dictionary, we adopt an off-the-shelf MT product[5] and the EDR J-E/E-J dictionary. We use the Google Web APIs[3] as a search engine.

[3] http://www.google.com/apis/

6.1 Overall evaluation of the E-J CLQA system

Table 1 shows the overall evaluation of the E-J CLQA system. We submitted four formal runs for the J-E task The metrics for the evaluation are "TOP1"[4], "MRR"[5], and "TOP5."[6]

Table 1. Overall evaluation of the E-J CLQA system (formal runs)

	TOP1	MRR	TOP5	TOP1 +U	MRR +U	TOP5 +U
MT	0.065	0.081	0.105	0.090	0.116	0.155
+G	0.075	0.088	0.105	0.100	0.125	0.160
+T	0.090	0.110	0.140	0.120	0.162	0.225
+G+T	0.125	0.141	0.160	0.155	0.190	0.240
JJQA	0.170	0.239	0.360	0.265	0.373	0.545

+U: Evaluation including unsupported answers
MT: Only an MT system and a bilingual dictionary are used.
+G: The translation information in Google's snippets is also used.
+T: Question types detected from English questions are also used.
JJQA: the Japanese-Japanese QA system. The upper bound.

As shown in Table 1, the proposed method ("+G+T") outperforms the baseline system ("MT"). The table also shows that both of the translation information obtained from the Web and the question types detected from English questions improve the accuracy of the E-J CLQA. Especially, the latter is more effective than the former. One of the reasons is that the Japanese QA system sometimes fails to detect question types from the machine-translated Japanese questions. The detailed analysis is described in Section 6.4.

6.2 Type-by-type evaluation of the E-J CLQA system

Table 2 shows the evaluation of the E-J CLQA system on a question-type-by-question-type basis. The labels Tn represents the number of questions whose answers are found within the top n answer candidates. The other notations are the same as Table 1. It should be noted that the question types in Table 2 are defined by the task organizers and are given in the data set of the formal run.

With respect to the types MONEY, TIME, and PERCENT, the methods without the question type detection in the source language fail to find correct answers. One of the reasons is the error in the translation of question as described later (see Section 6.4).

The proposed method (+G+T) sometimes fails to find answers with respect to the types NUMEX, ARTIFACT, and PERSON in comparison with the J-J QA. On the other hand, with respect to the types ORGANIZATION and LOCATION, the accuracy of the proposed method is almost same as the J-J QA.

6.3 Obtaining the translation information from the Web

Table 3 shows the overall accuracy of translation performed by each method for the Japanese and English proper noun candidates detected by the method described in Section 4.1. The token numbers of

[4] The ratio of correct answers to the first-ranked answer candidates.
[5] The average of the reciprocal rank of the highest correct answer.
[6] The ratio of the number of questions whose answers are found within the top five answer candidates.

Table 2. Type-by-type evaluation of the E-J CLQA system (formal runs)

	DATE/25				LOCATION/30			
	T1	T1 +U	T5	T5 +U	T1	T1 +U	T5	T5 +U
MT	6	7	7	9	2	2	3	3
+G	6	7	6	8	3	3	3	3
+T	6	7	7	10	2	3	3	4
+G+T	7	8	7	9	3	4	3	4
JJQA	9	12	15	18	2	6	5	9
	ORGANIZATION/26				PERSON/27			
	T1	T1 +U	T5	T5 +U	T1	T1 +U	T5	T5 +U
MT	2	4	3	5	2	3	3	5
+G	2	4	3	5	3	4	3	5
+T	2	4	3	5	2	3	3	5
+G+T	3	5	4	6	3	4	3	5
JJQA	3	7	8	13	6	9	9	15
	NUMEX/30				MONEY/20			
	T1	T1 +U	T5	T5 +U	T1	T1 +U	T5	T5 +U
MT	0	1	1	2	0	0	0	0
+G	0	1	2	4	0	0	0	0
+T	2	3	2	3	2	2	4	5
+G+T	3	4	3	5	3	3	4	5
JJQA	6	9	10	16	4	4	7	8
	ARTIFACT/18				TIME/14			
	T1	T1 +U	T5	T5 +U	T1	T1 +U	T5	T5 +U
MT	1	1	1	1	0	0	0	0
+G	1	1	2	3	0	0	0	0
+T	1	1	1	1	1	1	1	2
+G+T	1	1	2	3	1	1	1	2
JJQA	3	4	5	7	1	2	2	3
	PERCENT/10							
	T1	T1 +U	T5	T5 +U				
MT	0	0	0	0				
+G	0	0	0	0				
+T	0	0	1	2				
+G+T	1	1	1	2				
JJQA	0	0	4	5				

Japanese and English candidates are 112 and 97, respectively.

Table 3. The overall accuracy of translation of proper nouns

	MT	WEB	MT+WEB		
			OK	partially	N.G.
J.	36	57	71	10	31
E.	28	6	31	0	66

MT: MT only
WEB: Proposed method only
MT+WEB OK: Successfully translated by MT+WEB
MT+WEB partially: Partially translated by MT+WEB
MT+WEB N.G.: Translation is failed by MT+WEB

In order to evaluate each step in the translation, we define the following metrics.

- Metrics for the detection of the proper nouns to be translated.

$$\begin{aligned} R(\text{recall}) &= N_{detect}^{correct}/N_{detect}^{target} \\ P(\text{precision}) &= N_{detect}^{correct}/N_{detect}^{found} \end{aligned} \quad (2)$$

- Metrics for the translation of detected proper nouns.

$$\begin{aligned} H(\text{hit ratio}) &= N_{trans}^{hit}/N_{detect}^{found} \\ A(\text{accuracy}) &= N_{trans}^{correct}/N_{trans}^{hit} \end{aligned}$$

where a) N_{detect}^{target}, b)N_{detect}^{found}, c) $N_{detect}^{correct}$, d) N_{trans}^{hit}, and e) $N_{trans}^{correct}$ are a) the number of the word sequences that are not in the bilingual dictionary and should be detected as the targets of translation, b) the number of the sequences that the system detected as the targets, c) the number of the sequences that are correctly detected by the system, d) the number of the sequences for which the system can find certain translation candidates from the Web, and e) the number of the sequences for which the system finds correct translations from the Web.

Table 4. Translation of proper nouns by the Web

	R	P	H	A
E→J	61.4%	31.3%	67.0%	76.0%
J→E	7.14%	3.09%	20.6%	30.0%

R:recall, P:precision, H:hit ratio, A:accuracy

6.3.1 Extraction of Japanese translations

The average processing time was about nine seconds per word sequence[7]. As shown in Table 4, the system extracts many false positives in the detection phase. On the other hand, the translation phase works relatively well. The accuracy of translation is 76.0%.

6.3.2 Extraction of English translations

The average processing time was about seven seconds per word sequence. As shown in Table 4, the recall, the precision, the hit ratio, and the accuracy of translation are quite low. There is room to improve both the detection phases and the translation phases.

[7]The system is implemented by Perl 5 on a Linux machine (CPU: Pentium III 1GHz × 2, Memory: 2GB, OS: RedHat ver. 7.2.)

6.4 Question type detection in the source language

Table 5 shows the evaluation result of question type detection in the source language, i.e. in English. The labels MT+WEB and ENG represent the detection by the Japanese QA system using an MT and the Web and the detection by matching patterns in the source language, respectively.

Table 5. Accuracy of question type detection in the source language

200 questions in the Formal Run					
Q type	PER	LOC	ORG	DAT	TIM
# of Q	29	30	20	24	15
MT+WEB	100%	76.7%	15.0%	91.7%	6.7%
ENG	100%	83.3%	15.0%	95.8%	86.7%
Q type	INT	PRD	PCT	LEN	SPD
# of Q	0	1	10	3	2
MT+WEB	0%	0%	0%	0%	0%
ENG	0%	0%	70.0%	33.3%	0%
Q type	SPC	WGT	YEA	MNY	CNT
# of Q	1	1	3	20	18
MT+WEB	0%	0%	33.3%	0.5%	55.6%
ENG	0%	0%	100%	80.0%	94.1%
Q type	NUM	OTH	Total		
# of Q	0	23	200		
MT+WEB	0%	91.3%	55.6%		
ENG	0%	95.7%	79.5%		

PER:PERSON, LOC:LOCATION, ORG:ORGANIZATION, DAT:DATE, TIM:TIME, INT:INTERVAL, PRD:PERIOD, PCT:PERCENT, LEN:LENGTH, SPD:SPEED, SPC:SPACE, WGT:WEIGHT, YEA:YEAR, MNY:MONEY, CNT:COUNT& FREQ., NUM:OTHER NUM., OTH:OTHERS

In general, ENG is more accurate than MT+WEB. Especially, the types PERCENT,MONEY, and COUNT& FREQ. can hardly be detected by the method MT+WEB, but the method ENG achieves relatively higher accuracy. One of the reasons that the method MT+WEB fails in such cases is the mistranslation by the MT system.

7 Experimental Result 2: J-E Task

We use the sample test set (300 questions) and the formal run test set (200 questions) of NTCIR-5 CLQA1 J-E task as the evaluation. The parameters β and γ for the keyword translation are 0.5 and 0.7, respectively.

7.1 Answer mapping

Table 6 shows an empirical comparison of two types of answer mapping described in Section 5.2. The test set of the experiment is the sample test set. The system uses the top 150 documents that the search engine returned for each question.

As the table shows, both of systems with answer mapping (i.e. AM(S) and AM(R)) outperform the system (No AM) that just translate Japanese answer candidate to English words. With regard to answer mapping, the strict mapping is more effective in TOP 1 and MRR, while the relaxed mapping outperforms the strict mapping in TOP5. The results show that the relaxed mapping is pushing up the recall of answer, but it still needs some ways to filter out unnecessary mapped strings.

Table 6. Effectiveness of answer mapping

	TOP1	MRR	TOP5
No AM	0.045	0.076	0.120
AM(S)	0.090	0.131	0.190
AM(R)	0.085	0.130	0.200
	TOP1+U	MRR+U	TOP5+U
No AM	0.060	0.095	0.145
AM(S)	0.115	0.163	0.235
AM(R)	0.105	0.157	0.240

No AM: without the answer mapping
AM(S): with the strict answer mapping
AM(R): with the relaxed answer mapping

7.2 Translation of keywords

Table 7 shows the comparison of the keyword translation methods "EDR JE only" and "EDR JE+Prob." described in Section 5.1. In the experiment, the number of retrieved document is 150, and the answer mapping is the relaxed mapping. The test set of the experiment is the sample test set.

Table 7. Comparison of methods of keyword translation

	TOP1	MRR	TOP5
JE	0.085	0.130	0.200
JE+P	0.100	0.130	0.185
	TOP1+U	MRR+U	TOP5+U
JE	0.105	0.157	0.240
JE+P	0.130	0.170	0.235

JE: EDR JE only.
JE+P: EDR JE+Prob.

The table shows that the method "EDR JE+Prob." is more accurate than "EDR JE only."

7.3 Experimental result of the formal runs of J-E task

Table 8 shows the result of the formal runs. We submitted four formal runs for the J-E task.

Table 8. Overall evaluation of the J-E CLQA system (formal runs)

	TOP1	MRR	TOP5
JE1	0.030	0.046	0.070
JE2	0.085	0.115	0.165
JE3	0.080	0.110	0.160
JE4	0.045	0.069	0.115
	TOP1+U	MRR+U	TOP5+U
JE1	0.030	0.054	0.090
JE2	0.090	0.128	0.195
JE3	0.085	0.123	0.190
JE4	0.060	0.092	0.150

JE1: no answer mapping, EDR JE only
JE2: strict mapping, EDR JE only
JE3: relaxed mapping, EDR JE only
JE4: relaxed mapping, EDR JE+Prob.

The best one among four runs is "JE2," which is based on the strict answer mapping and the keyword translation only with the EDR JE dictionary (without the probability information).

7.4 Type-by-type evaluation of the J-E CLQA system

Table 9 shows the evaluation of the J-E CLQA system on a question-type-by-question-type basis. The notations are the same as Table 2 in Section 6.2.

Table 9. Type-by-type evaluation of the J-E CLQA system (formal runs)

	DATE/25				LOCATION/30			
	T1	T1 +U	T5	T5 +U	T1	T1 +U	T5	T5 +U
JE1	2	0	5	5	0	0	1	2
JE2	5	5	9	10	3	3	4	5
JE3	5	5	9	10	3	3	4	5
JE4	3	4	6	8	2	3	2	4
	ORGANIZATION/26				PERSON/27			
	T1	T1 +U	T5	T5 +U	T1	T1 +U	T5	T5 +U
JE1	0	0	0	1	2	0	0	0
JE2	1	1	2	4	1	2	4	4
JE3	1	1	2	4	1	2	4	4
JE4	0	0	1	3	0	0	1	3
	NUMEX/30				MONEY/20			
	T1	T1 +U	T5	T5 +U	T1	T1 +U	T5	T5 +U
JE1	0	0	0	0	1	1	2	2
JE2	3	3	4	4	2	2	2	3
JE3	2	2	3	3	2	2	2	3
JE4	0	0	2	2	2	2	2	2
	ARTIFACT/18				TIME/14			
	T1	T1 +U	T5	T5 +U	T1	T1 +U	T5	T5 +U
JE1	1	1	3	3	0	0	0	0
JE2	1	1	3	3	0	0	0	0
JE3	1	1	3	3	0	0	0	0
JE4	1	1	4	4	0	0	0	0
	PERCENT/10							
	T1	T1 +U	T5	T5 +U				
JE1	0	0	0	0				
JE2	1	1	4	4				
JE3	1	1	4	4				
JE4	1	1	2	2				

As shown by Run "JE1", the system without the answer mapping easily fails to find the answers in the question types DATE, PERCENT, and NUMEX, which are usually easy to extract with an NE recognizer. It is caused by the difference between the (English) answer in the English documents and the English translation of Japanese answers obtained by the Japanese QA system.

With regard to the type TIME, each system fails to find the answers in that type. One of the reasons is that the MT system cannot translate time expressions properly.

8 Conclusion

In this paper, we reported our proposals for CLQA using an existing mono-lingual QA system and some experimental results. With regard to the E-J task, we propose a method to translate proper nouns using the documents on the Web. Although we confirm that it works well for Japanese proper nouns, we have to revise the method for English proper nouns. We also introduce a pattern-matching-based question type detection in the source language in order to cope with translation errors. It is effective to improve the accuracy, but it fails to detect several question types. With regard to the J-E task, we adopt the bilingual-dictionary-based keyword translation, and in the translation we utilize the information about the classification of translation in the bilingual dictionary. We also introduce the answer mapping method to find answer expressions in English documents. Although the answer mapping is effective, but the absolute values of accuracy measures are very low. It is due to the difficulty in question answering using translated documents.

References

[1] K. Ahn, B. Alex, J. Bos, T. Dalmas, J. L. Leidner, and M. B. Smillie. Cross-lingual question answering with QED. In *Working Notes for the CLEF 2004 Workshop*, Sept. 2004.

[2] G. Bourdil, F. Elkateb, B. Grau, G. Illouz, L. Monceaux, I. Robba, and A. Vilnat. How to answer in English to questions asked in French: by exploiting results from several sources of information. In *Working Notes for the CLEF 2004 Workshop*, Sept. 2004.

[3] C. L. E. F. (CLEF). Cross language evaluation forum. http://clef.iei.pi.cnr.it/, 2005.

[4] IREX Committee, editor. *Proceedings of IREX workshop*. IREX Committee, 1999. (in Japanese).

[5] I. Japan. *Internet Hon'yaku-no Ousama (The king of Internet-translation Bilingual) Bilingual Version 5*, 2002.

[6] V. Jijkoun, G. Mishne, M. de Rijke, S. Schlobach, D. Ahn, and K. Müller. The University of Amsterdam at QA@CLEF 2004. In *Working Notes for the CLEF 2004 Workshop*, Sept. 2004.

[7] T. Mori. Japanese question-answering system using a* search and its improvement. *ACM Transactions on Asian Language Information Processing (TALIP)*, to appear. Special Issue for NTCIR-4.

[8] M. Murata, M. Utiyama, and H. Isahara. Question answering system using similarity-guided reasoning. SIG Notes 2000-NL-135, Information Processing Society of Japan, Jan. 2000.

[9] Y. Sasaki. A comprehensive learning approach to trainable QA: Toward CLQA. SIG Technical Reports 2004-NL-163, Information Processing Society of Japan, Sept. 2004. (in Japanese).

[10] K. Tsuji, S. Sato, and K. Kageura. Taiyaku jinmei-ni okeru hon'yaku·search-engine-no yukousei hyouka (evaluation of the effectiveness of transliteration and search engines in the translation of persons' names). In *Proceedings of the 11th annual meeting of the association for Natural Language Processing, Japan*, pages 352–355, Mar. 2005. (in Japanese).