A QA system that can answer any class of Japanese non-factoid questions and its application to CCLQA EN-JA task

— Yokohama National University at NTCIR-7 ACLIA CCLQA EN-JA —

Tatsunori MORI Takuya OKUBO Madoka ISHIOROSHI
Graduate School of Environment and Information Sciences
Yokohama National University
79-7 Tokiwadai, Hodogaya, Yokohama 240-8501, Japan
{mori,takuya04,ishioroshi}@forest.eis.ynu.ac.jp

Abstract

In this paper, we reported the evaluation results of our CCLQA system at NTCIR-7 ACLIA. We participated in the English-Japanese (EN-JA) cross-lingual task and the Japanese mono-lingual task. The system consists of a question translation module and a non-factoid-type Japanese question-answering system.

The question translation module was developed for NTCIR-6 CLQA, which is a combination of an off-the-shelf machine-translation product and a noun phrase translation module using web documents in order to compensate the insufficiencies in the bilingual dictionary of the MT product. With regard to the non-factoid-type Japanese question-answering system, we proposed a method of non-factoid question-answering that can uniformly deal with any class of Japanese non-factoid question by using a large number of example Q&A pairs.

Keywords: *EN-JA CCLQA, machine translation, noun phrase translation using the Web, Q&A pairs from a social Q&A website.*

1 Introduction

In this paper, we will report the evaluation results of our CCLQA system at NTCIR-7 ACLIA. We participated in the English-Japanese (EN-JA) cross-lingual task and the Japanese mono-lingual task. The system consists of a question translation module and a non-factoid-type Japanese question-answering system.

The question translation module was developed for NTCIR-6 CLQA. With regard to the matter of translation, many off-the-shelf machine-translation (MT) products are available in the market. Therefore we basically utilize one of off-the-shelf MT systems. However, in general, the quality of output of MT is not enough for the basis of CLQA. Especially, some proper nouns are not translated appropriately because of the OOV problem. The problem of OOV has very crucial impact on retrieval of question-related information from the text database. Thus, we introduced noun phrase translation using web documents in order to compensate the insufficiencies in the bilingual dictionary of the MT system.

With regard to the non-factoid-type Japanese question-answering system, we proposed a method of non-factoid (Web) question-answering that can uniformly deal with any class of Japanese non-factoid question by using a large number of example Q&A pairs. Instead of preparing classes of questions beforehand, the method retrieves already asked question examples similar to a submitted question from a set of Q&A pairs. Then, instead of preparing clue expressions for the writing style of answers according to each question class beforehand, it dynamically extracts clue expressions from the answer examples corresponding to the retrieved question examples. This clue expression information is combined with topical content information from the question to extract appropriate answer candidates. The method is suitable for not newspaper articles, but Web documents, because many of Web documents are written in colloquial styles, in which the Q&A pairs are also written.

2 Related studies

Table 1 shows typical classes of non-factoid questions and Japanese examples of the writing styles of questions and answers. Some fixed expressions are observed in both questions and answers according to the class of the question.

Answer candidates for such non-factoid questions tend to be descriptive expressions, which are relatively long and cover a series of sentences. As described by Han et al.[2] with regard to definitional question-answering, the appropriateness of such relatively long answer candidates can be estimated by the combination of, at least, the following two measures.

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. Measure 1 can be implemented as the content similarity between a given

Table 1. Typical classes of non-factoid questions

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

question and an answer candidate. In many previous studies, Measure 2 was estimated according to the application results of rules that detected certain writing styles.

Although the CCLQA of ACLIA is limited to handling four classes of questions, generally speaking, the classes of non-factoid questions are not well defined, and, therefore, 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 a different class. Therefore, a method is needed that does not depend on question classification and can deal with any class of questions uniformly.

Mizuno et al[3] proposed a method to realize Measure 2 without the classification of questions in Japanese non-factoid question answering. Using example Q&A pairs from a social Q&A website, it learns a binary classifier that judges whether or not the class of a given answer candidate is consistent with the class of a given question. By using this classifier, Measure 2 is realized without question classification. Because of the nature of the method, the length of the answer candidates should be predetermined as some text unit, like a paragraph. Therefore, the length of answer candidates is fixed and cannot be changed dynamically for a given question. With regard to the preparation of training data, negative examples should be artificially created by combining questions with answers from other questions.

Soricut et al.[7] also proposed an English non-factoid question answering system that does not need question classification. They introduced a statistical translation model between questions and the corresponding answers in order to bridge the lexical gap between questions and answers. A set of example Q&A pairs from FAQ sites on the Web are used for the estimation of the model. The model makes no distinction between the probability in terms of Measure 1 and that of Measure 2. Therefore, a large number of FAQs is supposed to be needed in order to guarantee the coverage of content words in questions. In this model, the length of the answers should be predetermined. Moreover, it requires a model that estimates the length of an answer from the length of the question.

3 System overview

The overview of the participating system is shown in Figure 1. Each question submitted by a user is translated into English question candidates in Part I shown in Figure 1. As shown in Part II of Figure 1 , the translated question candidates are separately processed by a non-factoid-type Japanese question-answering system. The Japanese question-answering system finds scored answer candidates from documents of information source by the steps that are explained in Part III of Figure 1.

We will describe each part of the system in the following sections.

4 Translation of questions

With regard to the matter of translation, many off-the-shelf machine-translation (MT) products are available in the market. Therefore we basically utilize one of off-the-shelf MT systems. However, in general, the quality of output of MT is not enough for the basis of CLQA. Especially, some proper nouns are not translated appropriately because of the OOV problem. The problem of OOV has very crucial impact on retrieval of question-related information from the text database.

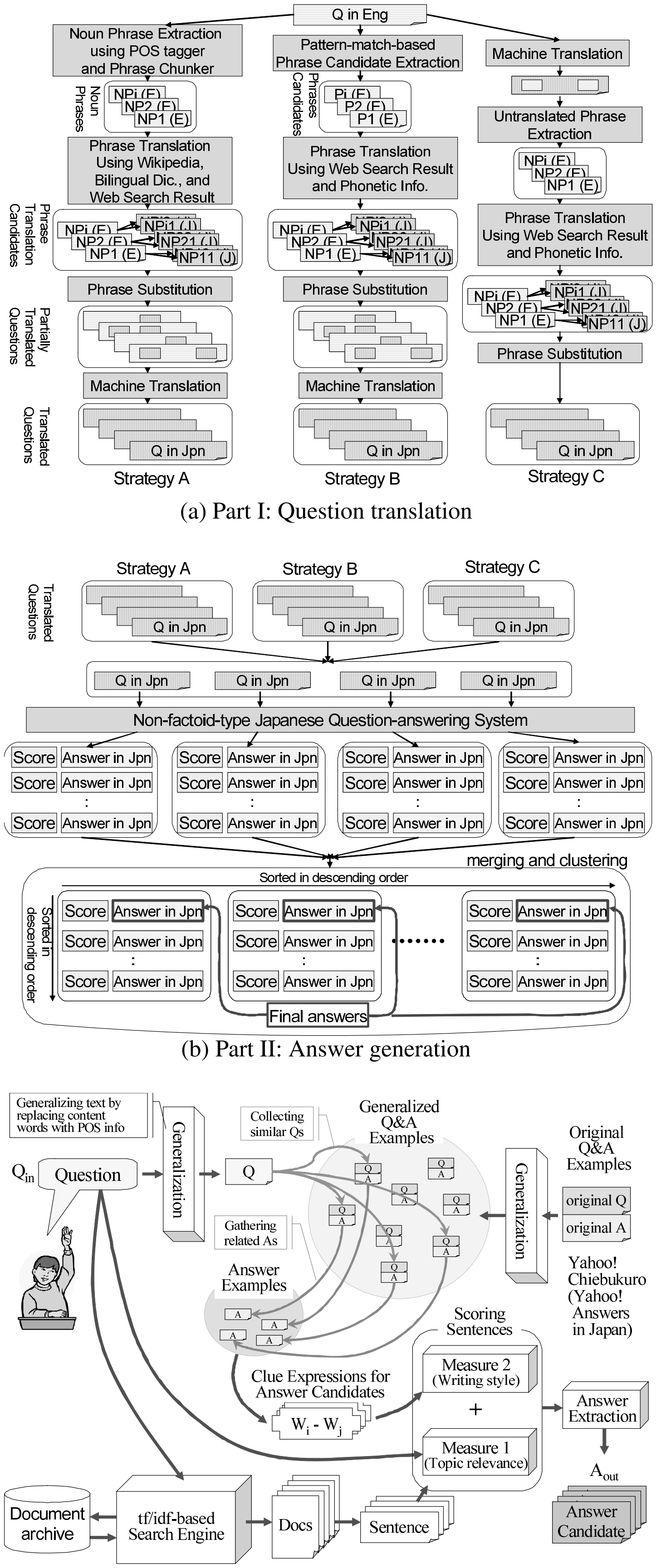
With regard to the treatment of OOV phrases in combination with an MT product, there are at least two types of approaches: the treatment in the pre-editing phase, and the treatment in the post-editing phase.

The latter may be easily performed independent of an MT system. However, the approach can only treat the phrases that are not translated by an MT system. Incorrect translations by an MT system will still remain in translated question sentences.

On the other hand, the former approach depends on the process of MT systems. Fortunately, some of off-the-shelf English-Japanese machine translation systems treat Japanese strings embedded in an English sentence as unknown noun phrases in the process of translation. The behavior can be used for *pre-translation*, which is one of techniques to utilize Translation Memory (TM)[1].

In the situation of EN-JA CLQA, the pre-translation module firstly identifies noun phrases and, then, try to translate them using some external translation resources. According to the result of phrase translation, the translated Japanese phrases are substituted for the original English phrases to generate partially translated question sentences. The question sentences are passed to the subsequent MT process. This pre-translation approach has the advantage that we can control the identification of phrase to be translated with external resources.

[1] In the pre-translation mode, TM system's proposals of translation are automatically inserted into source text. The produced hybrid text containing a mixture of source and target language elements is presented to human translators for further translation.


Q in Eng
Noun Phrase Extraction using POS tagger and Phrase Chunker
Pattern-match-based Phrase Candidate Extraction
Machine Translation
Noun Phrases
NPi (E)
NP2 (E)
NP1 (E)
Phrases Candidates
Pi (E)
P2 (E)
P1 (E)
Untranslated Phrase Extraction
Phrase Translation Using Wikipedia, Bilingual Dic., and Web Search Result
Phrase Translation Using Web Search Result and Phonetic Info.
Phrase Translation Candidates
NPi1 (J)
NP21 (J)
NP11 (J)
Phrase Substitution
Partially Translated Questions
Machine Translation
Translated Questions
Q in Jpn
Strategy A
Strategy B
Strategy C

(a) Part I: Question translation

Strategy A
Strategy B
Strategy C
Translated Questions
Q in Jpn
Non-factoid-type Japanese Question-answering System
Score
Answer in Jpn
merging and clustering
Sorted in descending order
Sorted in descending order
Final answers

(b) Part II: Answer generation

Generalizing text by replacing content words with POS info
Generalization
Collecting similar Qs
Generalized Q&A Examples
Original Q&A Examples
original Q
original A
Generalization
Q
A
Qin
Question
Gathering related As
Answer Examples
Yahoo! Chiebukuro (Yahoo! Answers in Japan)
Scoring Sentences
Measure 2 (Writing style)
+
Measure 1 (Topic relevance)
Clue Expressions for Answer Candidates
Wi - Wj
Answer Extraction
Aout
Answer Candidate
Document archive
tf/idf-based Search Engine
Docs
Sentence

(c) Part III: Non-factoid-type Japanese QA system in Part II

Figure 1. Overview of the proposed method

With regard to the phrase translation using external resources, there are several different approaches that are worth employing.

Therefore, as shown in Part I and II of Figure 1, we adopt a hybrid accroach that is a combination of pre-translation and post-translation along with several phrase translation methods.

4.1 Translation strategy A

Translation strategy A was introduced for NTCIR-6 CLQA, which is based on a pre-translation approach that utilizes 1) SVM-based noun phrase extraction, 2) phrase translation using Wikipedia, and 3) phrase translation using Web search results. See [6] for further detail.

4.2 Translation strategy B and C

Both of translation strategies B and C were developed by us for NTCIR-5 CLQA[4]. Each of them employs a translation strategy that searches for the loan words that are originally Japanese words, then, translates the loan words into the original Japanese words using the Web documents and the information of pronunciation. They also utilize a simple pattern-match-based method to find proper Japanese translations for English phrases using the Web documents.

With regard to the combination of phrase translation and MT, the translation strategy B is a pre-translation method. On the other hand, the translation strategy C is a post-translation method.

Please see [4] for further detail.

5 QA system that can answer any class of Japanese non-factoid questions

As shown in Table 1, non-factoid questions can be divided into several classes in terms of the content of their answers, for example, definition-type, why-type, how-type, and so on. Since clue expressions for answers are usually peculiar to the writing styles of each question class, many previous studies employed question classifiers to determine classes, and then applied one of several answer extraction methods, each of which was specific to a question class. While classes of factoid questions can be defined according to the categories in a thesaurus, classes of non-factoid questions are not well defined. With the exceptions of some typical classes like the definition-type, why-type, and how-type, it is difficult to distinguish and define all classes comprehensively.

In order to deal with this issue, we proposed a method to utilize a large number of example question-and-answer (Q&A) pairs from a social Q&A website[5]. Part III of Figure 1 shows an outline of our proposed method.

Instead of preparing classes of questions beforehand, this method retrieves already asked question examples that are similar to a submitted question from the set of Q&A pairs. Then, based on the writing style of the answers, it dynamically extracts clue expressions from the answer examples that correspond to the retrieved question examples. This clue expression information is combined with topical content information from the question to extract appropriate answer candidates. Note that we utilize the set of Q&A pairs, not to find answers from them, but to obtain clue expressions about the writing style of their answers.

This example-based method is expected to have the following advantages:

- It is free from the danger of question classification failure.
- Since extracted clue expressions are specific to not just a class of question but the submitted question itself, the clue expressions are more specialized and, therefore, expected to contribute to finding answer candidates that are more suitable to the question.

5.1 Obtaining clue expressions from Q&A examples

5.1.1 Q&A examples

We utilized a corpus of Q&A examples submitted to "Yahoo! Chiebukuro," which is a social Q&A website and the Japanese version of "Yahoo! answers." This corpus included about 3.1 million questions and 13.5 million answers that were contributed during the period from April 2004 to October 2005. Although each question had multiple answers, we utilized only the "best answers," which were selected as the best by the questioners. Hereafter, we use the term "Q&A pair" to refer to a pair consisting of a question and its best answer.

5.1.2 Generalizing texts in Q&A pairs

In order to extract only information about the writing style from examples for Measure 2, in this stage, we applied the following generalization to question texts and answer texts in the set of Q&A pairs. After word segmentation, the functional words, like interrogatives, postpositional particles, and so on, and a set of special content words described later are left as they are[2]. On the other hand, other words are replaced with their part-of-speech names. The set of special content words includes a set of Japanese content words that tend to be the foci of questions, like "*riyuu* (reason)," "*houhou* (method)," "*imi* (meaning)," "*chigai* (difference)," and so on. Verbs and adjuncts that appear with high frequency are also included in the set of special content words.

5.1.3 Generalizing question texts

From an examination of 30 Japanese sample questions from the evaluation workshop NTCIR-6 QAC[1], we found that a word 7-gram whose center word is the interrogative of the question generally provides

[2] More precisely, they are replaced with the strings that represent their pronunciations.

enough information to determine the class of question. Therefore, after the generalization described in Section 5.1.2, at this stage each example question was replaced with a 7-gram extracted from the question as a further generalization.

5.1.4 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 proposed method retrieves example Q&A pairs whose questions are similar to the submitted question from the viewpoint of writing style. As described before, a word 7-gram whose center word is an interrogative seems to give us enough context to determine the class of question. Therefore, we define the similarity between two questions as the similarity between the word 7-grams extracted from questions. According to the similarity, N-best example Q&A pairs are obtained by using an ordinary information retrieval technique.

5.1.5 Extracting clue expressions from answer examples

In this stage, clue expressions are extracted from the answers in the example Q&A pairs obtained in the stage described in Section 5.1.4. In this paper, we adopted a 2-gram as a clue expression unit because it is the smallest unit that can represent relations between words. Here, we assume 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 collected Q&A pairs.

As the measurement of the correlation, we adopted the χ^2 value shown in Equation (1) for the following two kinds of events for the answers from the entire set of example Q&A pairs:

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

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

$$\chi^2(b) = \frac{n}{|A| \cdot |\overline{A}| \cdot |B(b)| \cdot |\overline{B(b)}|} \cdot (|A \cap B(b)| \cdot |\overline{A} \cap \overline{B(b)}| - |\overline{A} \cap B| \cdot |A \cap \overline{B}|)^2 \quad (1)$$

where n is the total number of example Q&A pairs. The more correlated two events are, the larger the value of χ^2 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.

5.2 Question Answering using clue expressions obtained from Q&A examples

5.2.1 Extracting keywords from a question and obtaining their related words

From a question submitted by a user, 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 a 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 obtained snippets.
3. The weight value $T(w_j)$ defined as the following equation is calculated for each word w_j in snippets:

$$T(w_j) \quad = \quad \max_i \frac{freq(w_j, i)}{n_i} \quad (2)$$

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:

$$T(k) \quad = \quad \max_j T(w_j) \quad (3)$$

5.2.2 Grading sentences in retrieved documents

In this stage, each sentence in the retrieved documents is graded in terms of both Measure 1 and Measure 2. First, by using the method in Section 5.1.4, the system collects example Q&A pairs whose questions are similar to the submitted question in terms of writing style. Second, by using the method described in Section 5.1.5, it extracts a set of 2-grams as clue expressions from the answer examples of the example Q&A pairs and calculates the corresponding $\chi^2(b)$ value for each 2-gram b. Finally, the score of each sentence is calculated by using the following equation:

$$\text{Score}(S_i) = \frac{1}{\log(1 + \text{length}(S_i))} \cdot \left\{ \sum_{j=1}^{n} T(w_{i,j}) \right\}^{\gamma} \cdot \left\{ \sum_{k=1}^{m} \sqrt{\chi^2(b_{i,k})} \right\}^{1-\gamma} \quad (4)$$

where n is the number of different words in the sentence S_i, m is the number of different 2-grams in S_i, $w_{i,j}$ is the j-th word in sentence S_i, and $b_{i,k}$ is the k-th 2-gram in S_i. Since the terms $\sum_{j=1}^{n} T(w_{i,j})$ and $\sum_{k=1}^{m} \sqrt{\chi^2(b_{i,k})}$ in Equation (4) correspond to Measure 1 and Measure 2, respectively, the parameter γ is used to determine the mixture ratio of Measure 1 and Measure 2. The normalization term $\frac{1}{\log(1+\text{length}(S_i))}$ is introduced 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 correlate with answer examples). In order to reward longer sentences, a sentence length logarithm is adopted.

5.2.3 Extracting answer candidates

An outline of the extraction of answer candidates is shown in Figure 2. First, all sentences with maximal (not maximum) scores are selected from the retrieved documents . These play a role as the seeds of answer candidates. In this paper, an answer candidate corresponding to a seed is defined as the longest series of sentences that satisfies the following conditions: 1) the seed is in the series, and 2) every sentence in the series has a score greater than a threshold. The threshold is calculated seed by seed from a predetermined ratio in terms of score. For example, in Figure 2, the ratio is set as 50% of the maximal score. We define the score of an answer candidate as the maximal score.

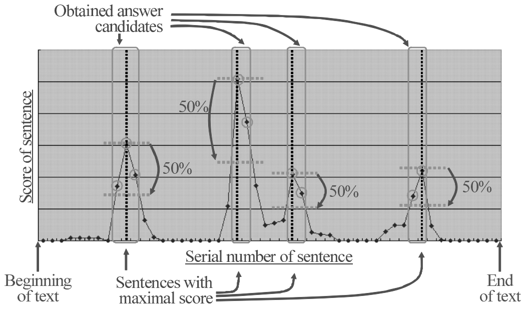

Obtained answer candidates
Score of sentence
50%
50%
50%
50%
50%
Serial number of sentence
Beginning of text
Sentences with maximal score
End of text

Figure 2. Generation of answer candidates

Since we may obtain similar answer candidates from different Web documents, we introduce the complete-link clustering of answer candidates in order to reduce the number of redundant answer candidates. For each cluster, the answer candidate with the highest score is obtained as a representative of the cluster.

The list of these representatives of the clusters with scores is the output of the non-factoid-type Japanese question-answering system. In the following experiments, the m-best representatives are adopted as the final output of the QA system.

6 Experiments

We conducted followings runs:

a The runs of the end-to-end EN-JA CCLQA and JA-JA monolingual QA using our proposed method.

b The runs with other IR results by IR4QA participants.

6.1 Setting of experiment

We employed the following tools, resources, and parameter settings in our experiment in NTCIR-7 ACLIA CCLQA.

Morphological analyzer : ChaSen[3].

Q&A corpus : Yahoo! Chiebukuro.

Web search engine for keyword expansion : API for the search engine of Yahoo! JAPAN.

Documents of information source : Mainichi Shimbun Newspaper articles 1998-2001.

Number of the answer candidates to be returned : Five (fixed).

6.2 Experimental results

6.2.1 The runs of the end-to-end EN-JA CCLQA and JA-JA monolingual QA using our proposed method

Figure 3 shows the evaluation of the run of the end-to-end EN-JA CCLQA using our proposed method in terms of F3.

According to the ‘Human-in-the-loop’ evaluation, the overall accuracy is far from sufficient. Especially, the questions about event lists are difficult for the system. However, the translation of keywords, e.g. named entities, in questions is not so poor. Figure 4 shows the accuracy of the translation. Here, the label “Correct (literally)” represents the situation that the English keywords are correctly translated into the identical words appeared in the Japanese questions. The label “Correct (semantically)” means that the translated words are not identical, but are semantically equivalent to the words in the Japanese questions. The label “Correct (partially)” means that the translated words share main parts with the words in Japanese questions.

Although more detailed analysis will be needed in order to make the reason clear, one possibility is that the set of Q&A examples we used does not contain this type of questions and effective clue expressions were not obtained. Especially, our method heavily depends on the existence of interrogatives in questions as described in Section 5.1.3. Figure 5 shows the relation between the average F3 values and the existence of interrogatives in the English and Japanese questions. Half of the topics do not contain interrogatives in both the English question and the Japanese question. The average F3 values for those topics are smaller than the topics whose questions have interrogatives. It should be noted that there are eleven topics in which the English questions have interrogatives but Japanese questions do not have. In those case, the average of F3 values for the EN-JA cross-lingual setting is much higher than one of the JA-JA monolingual setting.

With regard to automatic evaluations with POURPRE (BINARIZED), the answers for the questions about event lists seem to be overestimated in comparison with the ‘Human-in-the-loop.’ Figure 6 shows the Kendall tau rank correlation coefficient between automatic evaluations with POURPRE (BINARIZED) and the ‘Human-in-the-loop.’ It may suggest that unigram is not always sufficient as the unit of information, when we evaluate QA systems.

[3] http://chasen-legacy.sourceforge.jp/

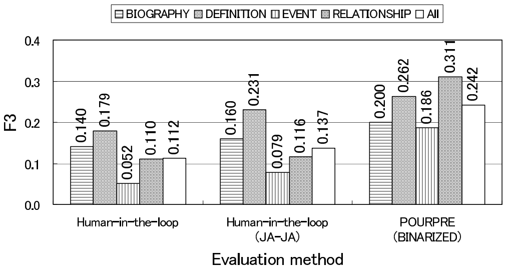


Figure 3. Evaluation of EN-JA CCLQA

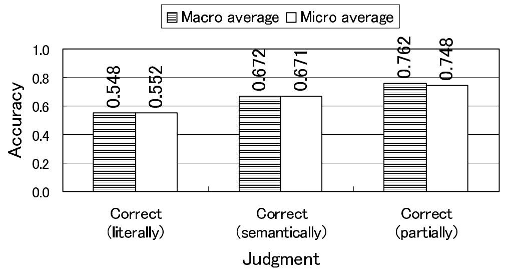


Figure 4. Accuracy of translation of keywords

6.2.2 The runs with other IR results by IR4QA participants

Figure 7 shows the evaluation of the runs with other IR results by IR4QA participants. The all settings are same as our CCLQA, but the IR results generated by other IR4QA participants were used instead of our IR result. This figure shows the impact of difference of CLIR method in the context of the end-to-end CCLQA task with our QA system.

Note that our IR module that is employed in the QA system is very simple one. It is a monolingual system and is based on the tf·idf method for term weighting and the vector space model for the calculation of similarity. Therefore, the main part of our contribution to CLIR lies in the question translation shown in Figure 1. It also should be noted that the questions translated by our method are also used as the inputs for the QA system in the runs with other IR results. As shown in Figure 7, the IR runs named CMUJAV-EN-JA-0[0-5]-T improve the accuracy of QA result in terms of automatic evaluation, especially for DEFINITION-type questions. On the other hand, the accuracy is degraded when we use the IR runs named TA-EN-JA-0[1-3]-.*.

Table 2 shows the result of statistical test of significant superiority in terms of F3 of automatic evaluation by using Wilcoxon matched pairs signed rank sum test. The runs named CMUJAV-EN-JA-0[0-5]-T are significantly improves the accuracy of answers for DEFINITION-type questions.

7 Conclusion

In this paper, we reported the evaluation results of our CCLQA system at NTCIR-7 ACLIA. The system consists of a question translation module and a non-factoid-type Japanese question-answering system. The question translation module was developed for NTCIR-6 CLQA, which is a combination of an off-the-shelf machine-translation product and a noun phrase translation module using web documents. With regard to the question-answering system, we proposed a method of non-factoid question-answering for Web documents that can uniformly deal with any class of Japanese non-factoid question by using a large number of example Q&A pairs. Although the combination can deal with cross-lingual complex questions for newspaper articles in ACLIA to some extent, the overall accuracy is far from sufficient.

Acknowledgment

We would like to thank Yahoo Japan Corporation and The National Institute of Informatics who provide us the data set of Yahoo! Chiebukuro. We are also grateful to people who manage the NTCIR workshops. We would like to especially thank Mainichi Shimbun for permitting us to use the documents for research.

This study was partially supported by Grant-in-Aid for Scientific Research on Priority Areas (No.19024033) from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

References

[1] J. Fukumoto, T. Kato, F. Masui, and T. Mori. An Overview of the 4th Question Answering Challenge (QAC-4) at NTCIR Workshop 6. In *Proceedings of the Sixth NTCIR Workshop Meeting*, pages 433–440, 5 2007.

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

[3] J. Mizuno, T. Akiba, A. Fujii, and K. Itou. Non-factoid Question Answering Experiments at NTCIR-6: Towards Answer Type Detection for Realworld Questions. In *Proceedings of the Sixth NTCIR Workshop*, pages 487–492, May 2007.

[4] T. Mori and M. Kawagishi. A Method of Cross Language Question-Answering Based on Machine Translation and Transliteration: Yokohama National University at NTCIR-5 CLQA1. In *Proceedings of the Fifth NTCIR Workshop Meeting*, pages 215–222, 12 2005.

[5] T. Mori, M. Sato, and M. Ishioroshi. Answering any class of japanese non-factoid question by using the Web and example Q&A pairs from a social Q&A website. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence (WI-2008)*, Dec. 2008. (to appear).

[6] T. Mori and K. Takahashi. A Method of Cross-Lingual Question-Answering Based on Machine Translation and Noun Phrase Translation using Web documents — Yokohama National University at NTCIR-6 CLQA —. In *Proceedings of the Sixth NTCIR Workshop Meeting*, pages 182–189, 5 2007.

[7] R. Soricut and E. Brill. Automatic Question Answering Using the Web: Beyond the Factoid. *Journal of Information Retrieval - Special Issue on Web Information Retrieval*, 9(2):191–206, Mar. 2006.

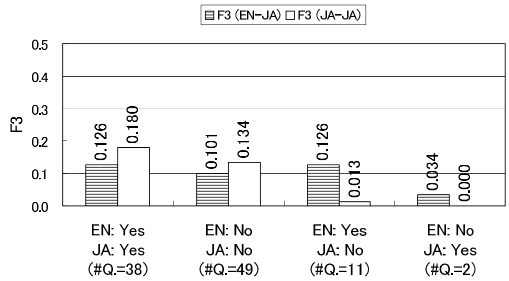


Figure 5. Average F3 values and the existence of interrogatives in questions

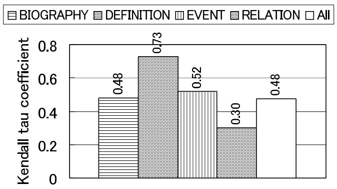


Figure 6. Kendall tau rank correlation coefficient of the automatic evaluations with the 'Human-in-the-loop' evaluation

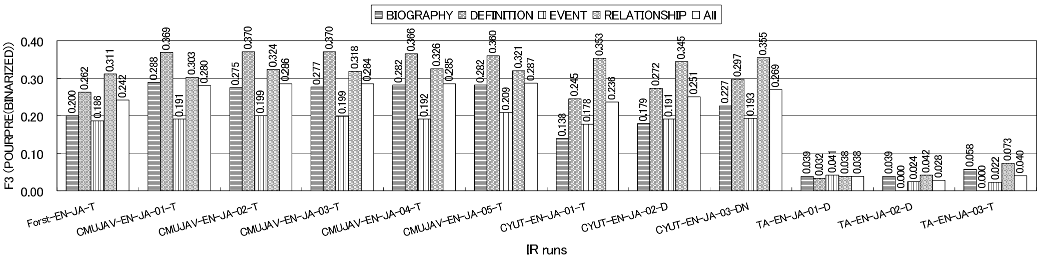


Figure 7. Automatic evaluation (POURPRE (BINARIZED)) with other IR results

Table 2. Statistical test of significant superiority in terms of F3 by Wilcoxon matched pairs signed rank sum test

		CMUJAV-EN-JA-01-T	CMUJAV-EN-JA-02-T	CMUJAV-EN-JA-03-T	CMUJAV-EN-JA-04-T	CMUJAV-EN-JA-05-T	CYUT-EN-JA-01-T	CYUT-EN-JA-02-D	CYUT-EN-JA-03-DN	TA-EN-JA-01-D	TA-EN-JA-02-D	TA-EN-JA-03-T
Forst-EN-JA-T	BIOGRAPHY									>>>	>>>	>>>
	DEFINITION		<<							>>>	>>>	>>>
	EVENT									>>>	>>>	>>>
	RELATION									>>>	>>>	>>>
	All		<<			<<				>>>	>>>	>>>
CMUJAV-EN-JA-01-T	BIOGRAPHY						>>			>>>	>>>	>>>
	DEFINITION						>>			>>>	>>>	>>>
	EVENT									>>>	>>>	>>>
	RELATION									>>>	>>>	>>>
	All									>>>	>>>	>>>
CMUJAV-EN-JA-02-T	BIOGRAPHY									>>>	>>>	>>>
	DEFINITION						>>	>>		>>>	>>>	>>>
	EVENT									>>>	>>>	>>>
	RELATION									>>>	>>>	>>>
	All						>>			>>>	>>>	>>>
CMUJAV-EN-JA-03-T	BIOGRAPHY									>>>	>>>	>>>
	DEFINITION						>>	>>		>>>	>>>	>>>
	EVENT									>>>	>>>	>>>
	RELATION									>>>	>>>	>>>
	All						>>			>>>	>>>	>>>
CMUJAV-EN-JA-04-T	BIOGRAPHY						>>			>>>	>>>	>>>
	DEFINITION						>>	>>		>>>	>>>	>>>
	EVENT									>>>	>>>	>>>
	RELATION									>>>	>>>	>>>
	All									>>>	>>>	>>>
CMUJAV-EN-JA-05-T	BIOGRAPHY						>>			>>>	>>>	>>>
	DEFINITION						>>	>>		>>>	>>>	>>>
	EVENT									>>>	>>>	>>>
	RELATION									>>>	>>>	>>>
	All						>>			>>>	>>>	>>>
CYUT-EN-JA-01-T	BIOGRAPHY								<<			
	DEFINITION									>>>	>>>	>>>
	EVENT									>>>	>>>	>>>
	RELATION									>>>	>>>	>>>
	All									>>>	>>>	>>>
CYUT-EN-JA-02-D	BIOGRAPHY									>>>	>>>	>>>
	DEFINITION									>>>	>>>	>>>
	EVENT									>>>	>>>	>>>
	RELATION									>>>	>>>	>>>
	All									>>>	>>>	>>>
CYUT-EN-JA-03-DN	BIOGRAPHY									>>>	>>>	>>>
	DEFINITION									>>>	>>>	>>>
	EVENT									>>>	>>>	>>>
	RELATION									>>>	>>>	>>>
	All									>>>	>>>	>>>
TA-EN-JA-01-D	BIOGRAPHY											
	DEFINITION											
	EVENT										<<	
	RELATION											
	All											
TA-EN-JA-02-D	BIOGRAPHY											
	DEFINITION											
	EVENT											
	RELATION											
	All											

>>, >>> : mean that the run at the row is significantly superior to the run at the column, $p < 0.05$ or $p < 0.01$, respectively.
<<,<<< : mean that the run at the column is significantly superior to the run at the row, $p < 0.05$ or $p < 0.01$, respectively.