A transitive translation for Indonesian-Japanese CLQA

Ayu Purwarianti,†1 Masatoshi Tsuchiya†2
and Seiichi Nakagawa†1

Abstract This paper investigates a Cross Language Question Answering for a language pair with limited resources. In order to solve the limited translation resource problem, we use a transitive translation with English as the pivot language. To select the best translation, we use a mutual information score calculated from the corpus of the target language (e.g. Japanese). As an additional advantage, we build the CLQA using some machine learning based modules. By this, the system can be easily adapted into other language pair. In the experiments, we use test set of NTCIR 2005 CLQA1 task, consists of 200 questions. We also compare the transitive translation result with a direct translation result which employs a middle size Indonesian-Japanese dictionary.

Keyword Cross language question answering, Indonesian-Japanese translation, transitive translation, mutual information, text chunking

1. Introduction and Related Works

CLQA (Cross Language Question Answering) systems have become an interesting research area. CLEF (Cross Language Evaluation Forum) and NTCIR (NII Test Collection for IR Systems) have provided CLQA task since several years ago. In CLEF, other than European language, since 2004 an Indonesian-English CLQA task has been conducted. In NTCIR, the languages involved are English, Japanese, Chinese, and Korean. In our knowledge, so far, no Indonesian-Japanese CLQA system or data are available.

A translation process made a CLQA system becomes a more difficult task than a monolingual QA system because different expressions are used in the questions and in the correct article. Here, the translation quality becomes an important factor in achieving an accurate QA system. Thus, the quality of the machine translation or dictionary used is very important.

For Indonesian-Japanese pair, there is a middle size dictionary available (Sanggar Bahasa Indonesia Proyek (2000), 14,823 vocabulary size). Other than this dictionary, no other Indonesian-Japanese translation resource (dictionary, machine translation or multilingual corpora) is available. In other side, for Indonesian-English, there are some translation resources available such as Indonesian-English KEBI dictionary (29,054 vocabulary size), Kataku Indonesian-English machine translation (http://www.toggletext.com/kataku_trial.php), and Transtool Indonesian-English machine translation (http://www.geocities.com/cdpenerjemah). This condition is common for a limited resource language where the available translation resource is for a pair of that certain language and English as the most known language. This phenomenon motivates researchers to do transitive translation for cross language system. Ballesteros[2)] translated Spanish queries into French with English as the interlingua. They used Collins Spanish-English and English-French dictionaries. Gollins and Sanderson[5)] translated Germany queries into English using two pivot languages (Spanish and Dutch). In our research, we use English as the interlingua for our Indonesian-Japanese question translation. Here we choose to employ bilingual dictionary as the most available translation resource rather than machine translation or multilingual corpora.

A transitive translation gives more translation alternatives than a direct translation. In order to select the most appropriate translation, a monolingual corpus can be used. Ballesteros and Croft[3)] used an English corpus to select adequate English translations based

†1 Department of Information and Computer Sciences, Toyohashi University of Technology
†2 Information and Media Center, Toyohashi University of Technology

on Spanish-English translation and analyzed the co-occurrence frequencies to disambiguate phrase translations. Gao[4] used a Chinese corpus to select the best English-Chinese translation set. They modified the EMMI weighting measure (a measure to estimate term similarity) by adding word distance factor. Qu[8] selected the best Spanish-English and Chinese-English translation using English corpus. The coherence score calculation was based on web page count, retrieval score and mutual information score. Adriani[1] translated Indonesian into English and used an English monolingual corpus to select the best translation, employing a term similarity score based on the Dice similarity coefficient. Here, we select the best Japanese translation using mutual information score and TF $\times$ IDF score.

Similar with a common approach, our CLQA system consists of a question analyzer (Indonesian), question keyword translation (Indonesian-Japanese), passage retrieval (Japanese) and answer finder (Japanese). The main different thing with the common CLQA system is that we do not use a named entity tagger to process the target documents, instead we try to locate the answer directly from POS-tagged target documents by using a text chunking approach. A similar approach was conducted by Sasaki[9] for English-Japanese CLQA. The method is called Extended QBTE (Question-Biased Term Extraction). Different with the common approach that calculates the similarity between an EAT (Expected Answer Type) and named entities in the target document, this approach tries to eliminate the question classification (to get the EAT) and named entity tagger processes. This method extracts the answer by classifying each word in the document into one of 3 classes (B, I or O). B class means that the document word is the first word in the answers, I means that the word is part of the answer and O means that the word is not part of the answer. This answer classification was done using a maximum entropy algorithm with features taken from Chasen that includes four information POS. The accuracy score obtained in the NTCIR 2005 CLQA was low, as there was only 1 correct answer among 200 test questions for the English-Japanese CLQA. We try to modify this method in our answer finder by using the question analyzer result having the EAT.

The best result on English-Japanese CLQA 2005 was achieved by FORST[6] which used machine translation software and web search to translate proper nouns. The accuracy result was 15.5% accuracy (31 correct answers as Top1 among 200 questions). In the answer finder they used a matching score of each morpheme in the retrieved passages with the question keywords and EAT.

The rest of this paper is organized as follows. In Section 2, we describe the modules used in our Indonesian-Japanese CLQA system. Section 3 shows the experimental results. Section 4 is the conclusion.

2. Modules in Indonesian-Japanese CLQA

There are some approaches can be used for a CLQA system. Among others are:

(1) The source keyword question is translated into target language. The question classifier is conducted for the original question. The passage retriever and answer finder are done for the target language.

(2) Use a pivot language. The source questions and target documents are translated into the pivot language. All the processes are done in the pivot language. However, there is a big risk that the system could not get advantage from the sentence pattern to locate the answer due to the document translation process.

(3) Similar with the second schema but the passage retriever is done on the target passages. The translation is done on the retrieved target passages.

(4) Another alternative is to retrieve the pivot language passages and then translate the passages into target language and do the answer finder in the target language.

There are other alternative approaches such as the modification on the question analyzer such as employed in the literature 6) and 10). We choose to use the first approach and employ the transitive translation using bilingual dictionaries in order to achieve the aim of easily ported system. The complete schema of our Indonesian-Japanese CLQA is shown in **Fig. 1**.

We do the Indonesian-Japanese translation not on the question sentence but on the keywords extracted from the question sentence[7]. First, an Indonesian question sentence is processed by a question analyzer. The results are then processed by the translation module to get the Japanese keywords. These Japanese keywords are used to retrieve Japanese passages. In the last step, a Japanese answer finder classifies the answer candidates from Japanese passages using Japanese keywords, Japanese question main word and the output of the question analyzer.

2.1 Indonesian Question Analyzer

This system consists of two main components: a question shallow parser and a question classifier. The question shallow parser aims to identify question information using some simple rules. The question information includes question keywords, question main word, interrogative word, and phrase information. The question classifier aims to extract the EAT (expected answer type) using an SVM algorithm. It classifies a question into one of the following classes: date, location, person, organization, name and quantity. The main input data used in the question classifier are the question information resulted by the question shallow parser. **Fig. 2** shows a result example of the question analyzer.

To enhance the quality of question classifier, we add a statistical value for the question main word, called as a bi-gram frequency score. This score is the frequency score of a bi-gram pair between the question main word and each word included in the predefined word list. There are 6 predefined word list, one for each class of "date", "location", "name", "organization", "person", and "quantity". For example, as shown in Fig. 2, for the word "partai", the biggest bi-gram frequency score is for "organization" class: 0.0789 and 22, where 0.0789 shows the frequency of the predefined word list of "organization" class with the word "partai" normalized with other classes and 22 is the number of unique words in the predefined word list of "organization" class which has "partai" as its bi-gram pair, respectively.

We also add a WordNet distance information as another additional attribute. The WordNet distance means a distance between the question main word (translated into English) and

Question: Pemerintahan koalisi Perdana Menteri Obuchi terdiri atas LDP, Komei, dan partai apa lagi? (Prime Minister Obuchi's coalition government consists of LDP, Komei, and what other party?)
Question interrogative: apa (what)
Question main word: partai (party)
Phrase information: NP-POST
Question keyword: pemerintahan (government), koalisi (coalition), Perdana Menteri (prime minister), Obuchi, LDP, Komei
Bi-gram frequency score: date(0,0), location(0.0101,11), name(0.0040,9), organization (0.0789,22), person(0.0005,5), quantity(0,0)
WordNet distance score: all are zero except for event(0.1724), group(0.4828), and person(0.3448)
Question classifier result: organization

Fig. 2 Example on the Question Analyzer's Output

some specified WordNet synsets (taken from the 25 noun lexicographer files in WordNet). Fig. 2 shows that the WordNet distance scores for the word "partai" ("party") are 0.1724 for event, 0.4828 for group and 0.3448 for person. It means the word "partai" has the biggest tendency to be defined as group.

2.2 Indonesian-Japanese Keyword Translator

Based on the observations on the collected Indonesian questions, there are three word types used in the Indonesian questions: native Indonesian words, English words (such as "barrel", "cherry", etc) and transformed English words (such as "prefektur" from "prefecture", "agensi" from "agency", etc). For native Indonesian words, we use two dictionaries: Indonesian-English (KEBI; 29,047 entries) and English-Japanese (Eijirou; 556,237 entries). If a word is not available in the Indonesian-English dictionary, we assumed it as an OOV word and will be treated as the second or third word type.

For the second word type, we use the English-Japanese dictionary, a Japanese proper name dictionary, a Japanese corpus and a katakana/hiragana transliteration module. The katakana information provided in the Japanese proper name dictionary is transliterated into alphabet. The word matching is done on this alphabet information.

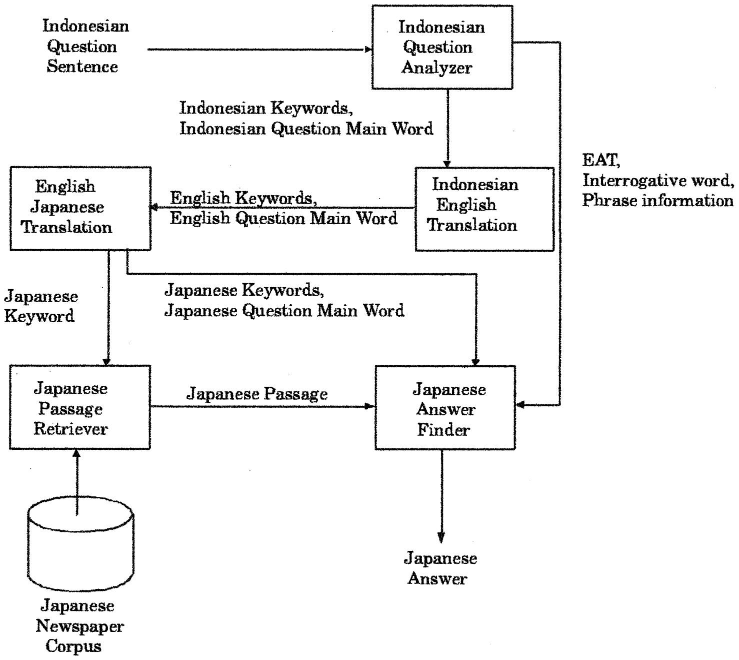

Fig. 1 System Architecture of Indonesian-Japanese CLQA with Transitive Translation

This is also done for the Japanese corpus. First, the Japanese corpus is morphologically analyzed by Chasen (http://chasen-legacy.sourceforge.jp). The katakana information resulted by Chasen is transliterated into alphabet. As for the third word type, we use some transformation rules to transform the Indonesian words into English words. The resulted English words are then assumed as the second word type.

The example on the translation result is shown in **Fig. 3**. The first and second rows are the example of the first word type. The third, fourth and fifth rows are the example of the second word type. The last row is the example of the third word type. The Indonesian word "prefektur" is not recorded in the dictionary, it is a borrowed word from the word "prefecture". To translate it into Japanese, we change the word into English first using some transformation rules and then we use English-Japanese dictionary to have the Japanese translation.

2.3 Japanese Document Retrieval

We use GETA (http://geta.ex.nii.ac.jp/) as our generic information retrieval engine. It retrieves some Japanese documents from a keyword set by using IDF, TF or TF × IDF score. The Japanese translation results are joined into one query and inputted into GETA to get the relevant Japanese documents.

Translation:
- pemerintahan koalisi (coallision government): 連合政権
- perdana menteri (prime minister): 首相, 総理大臣
- Obuchi: 小渕, 尾駮, 小淵, 大渕
- LDP: 自民党, 自由民主党
- Komei: こうめい, 光明, 公明, 広明, 昂明, 港明, 高明
- prefektur (prefecture) : 県、県庁、府、知事、長官

Fig. 3 Example on the Question Translation's Output

Because translation results using bilingual translation contain many Japanese translations, we try to filter the translation by using mutual information and TF × IDF score. First, all combinations of the keyword translation sets are ranked based on their mutual information

Document word: 自民党
POS information (Chasen): 名詞, 固有名詞, 組織
Similarity score: 0
5 preceding words: に, よる, 連立 (similarity score is 1), 政権, は
5 succeeding words: 執行, 部, が, 小渕 (similarity score is 1), 前

Fig. 4 Example of Document Feature for the Answer Finder Module

score. Each set of top 5 mutual information score is used to retrieve the relevant documents. In the final phase, we select documents within 100 highest TF × IDF score from all relevant documents resulted from the queries of top 5 mutual information score.

2.4 Japanese Answer Finder

Our Japanese answer finder locates the answer candidates using a text chunking approach with machine learning algorithm. Here, the document feature is directly matched with the question feature. Each word in the retrieved passages is classified into B (first word of the answer candidate) or I (part of the answer candidate) or O (not part of the answer candidate).

The features used for the classification include the document feature, question feature, EAT information and similarity score. The document feature includes POS information (four POS information resulted by Chasen), and the lexical form. The question feature includes the question shallow parser result and the translated question main word. The similarity score is the similarity score between the document word and the question keywords. An example of document feature for a document word "自民党" with question "Prime Minister Obuchi's coalition government consists of LDP, Komei, and what other party?" is shown in **Fig. 4**. For one document word, there are n preceding words and n succeeding words. In our experiment, we found that n=5 is the best option, such as shown in Fig. 4.

3. Experiments

In order to gain an adequate number of training data for the question classifier and the answer finder modules, we collected our Indonesian-Japanese CLQA data. So far, we have 2,837 Indonesian questions and 1,903 answer tagged Japanese passages. As for the test data, we translated 200 English questions test data of the NTCIR 2005 CLQA1 Task into Indonesian. The 200 Indonesian test questions contain 625 common nouns and 294 proper nouns. The Japanese corpus is the Yomiuri Newspaper corpus years 2000-2001 (658,719 articles).

Table 1 OOV Rates of Proper Noun and Common Noun Translation

Description	Proper Noun	Common Noun
Direct Translation	12.9%	17.2%
Transitive Translation	13.9%	9.4%

In the question classifier, we used an SVM algorithm from WEKA software (http://www.cs.waikato.ac.nz/nl/weka) with a linear kernel and the "string to word vector" function. We used 10-fold cross validation for the accuracy calculation. We tried some feature combinations of the bag-of-words feature, the shallow parser result, the bi-gram frequency score and the WordNet distance score. The highest result was 96.0% accuracy, achieved by the combination of the shallow parser result, bi-gram frequency score and WordNet distance score.

In the question translation, we compared our transitive translation with the direct translation. For the transitive translation, we used Indonesian-English KEBI dictionary (http://nlp.aia.bppt.go.id/kebi/, 29,054 vocabulary size) and English-Japanese Eijirou dictionary (556,237 vocabulary size). For the direct translation, we used Indonesian-Japanese Sanggar Bahasa Indonesia Proyek (2000) dictionary with 14,823 vocabulary size. As the additional resources to translate the OOV words, we used Japanese proper name dictionary and list of romanized words of Yomiuri Newspaper corpus. The OOV rates in query sentence (test data) are shown in **Table 1**. OOV words are the words that could not be translated by our translation module. In total for proper noun and common noun, the OOV rates were about 15.2% and 11.5% for the direct translation (middle size Indonesian-Japanese dictionary) and transitive translation (Indonesian-English and English-Japanese dictionary), respectively.

Further, we evaluated the translation result (direct and transitive translation) based on the

Table 2 Document Retrieval's Experimental Results on Indonesian-Japanese CLQA

Description	Recall	Precision
Direct Translation		
No filtering	35.0%	2.26%
1st MI score	37.5%	2.56%
2nd MI score	35.0%	2.47%
3rd MI score	36.0%	2.41%
4th MI score	36.0%	2.40%
5th MI score	37.0%	2.19%
Top 3 MI	36.0%	2.50%
Top 5 MI	35.5%	2.35%
Top 10 MI	35.5%	2.35%
MI-TF $\times$ IDF	35.5%	2.37%
Transitive Translation		
No filtering	30.0%	1.56%
1st MI score	34.5%	1.64%
2nd MI score	36.0%	1.71%
3rd MI score	36.0%	1.49%
4th MI score	35.0%	1.71%
5th MI score	35.0%	1.84%
Top 3 MI	32.0%	1.62%
Top 5 MI	35.5%	1.76%
Top 10 MI	38.5%	1.89%
MI-TF $\times$ IDF	35.5%	2.37%
Keywords of Japanese monolingual queries		
No filtering	70.0%	5.46%

Table 3 Answer Finder's Experimental Results for Correct Documents (in %)

Desc	Top1	Top5	TopN	MRR
Use transitive translation to calculate word similarity				
baseline	24.5	33.5	41.5	30.2
+word distance	27.5	40.0	42.5	32.8
Use keywords of Japanese queries to calculate word similarity				
baseline	25.0	38.0	41.0	30.2
+word distance	26.5	43.5	46.5	33.5

document retrieval score because our translation module aims to have a good retrieval results. The document retrieval results of these two methods along with the effect of the mutual information filtering process are shown in **Table 2**. Table 2 shows two evaluation measures: precision and recall. Precision shows the average ratio of relevant document. A relevant document is a document that contains a correct answer without considering any available supporting evidence. Recall refers to the number of questions that might have a correct answer in the retrieved documents. n-th MI score means the input query is the keyword set with the n-th rank of MI score. Top n MI means that the input query consists of keywords from the 1st rank until n-th rank of MI score. MI-TF $\times$ IDF is the combination of Mutual Information score and the TF $\times$ IDF score as explained in 2.3.

In the keyword translation, even though the number of OOV for common noun resulted by the direct translation is much larger than the transitive translation, but in general the direct translation has better retrieval performance than the transitive translation (higher precision score for all schemas and higher recall score for almost all schemas). It shows that the important keyword in the document retrieval mostly are the proper noun (number of OOV proper noun of direct translation is lower than the transitive translation).

Table 2 also shows that without the combination of TF $\times$ IDF and mutual information filtering, the transitive translation result will have lower recall score than the direct translation. It indicates that the combined filtering is effective for transitive translation result because it is able to decrease the incorrect Japanese translations. For the direct translation, the combined filtering is not effective because the number of Japanese translations is much fewer than the transitive translation. Table 2 also shows that our translation only achieved 50% accuracy compared to the oracle experiment (last row, the document retrieval using keywords extracted from Japanese monolingual queries).

In the answer finder module, we used Yamcha (http://chasen.org/ taku/software/yamcha) with default configuration as the SVM based text chunking software. To evaluate our answer finder module, we conducted the answer finder experiments for the correct passages. The result is shown in **Table 3**. The evaluation scores are Top1 (correct rate of the top 1 answers), Top5 (rate of at least one correct answer included in the top 5 answers), TopN (rate of at least one correct answer retrieved in the found answers) and MRR (Mean Reciprocal Rank, the average reciprocal rank (1/n) of the highest rank n of a correct answer for each question). Baseline means that we use features mentioned in 2.4. We tried to add an additional feature called the word distance which shows the distance between the current document word and other document word that equals to question keywords.

To see the effect of our transitive translation in the answer finder feature, we conducted two kinds of experiments for the oracle correct doc-

Table 4 Answer Finder's Experimental Results for Retrieved Documents (in %)
R: document retrieval's recall score
T: text chunking score

Desc	Top1	Top5	TopN	MRR
Direct Translation (baseline answer finder)				
0.3R+0.7T	1.5	5.5	14.0	3.8
0.5R+0.5T	2.0	6.0	14.0	4.2
0.7R+0.3T	2.5	7.0	14.0	4.9
Direct Translation (+word distance feature)				
0.3R+0.7T	2.0	7.0	15.0	4.4
0.5R+0.5T	2.0	7.5	15.0	4.4
0.7R+0.3T	3.0	6.5	15.0	5.2
Transitive Translation (baseline answer finder)				
0.3R+0.7T	1.5	6.0	19.0	3.9
0.5R+0.5T	1.5	6.5	19.0	3.9
0.7R+0.3T	2.0	7.0	19.0	4.4
Transitive Translation (+word distance feature)				
0.3R+0.7T	2.0	6.0	19.5	4.1
0.5R+0.5T	2.0	6.0	19.5	4.2
0.7R+0.3T	2.0	6.0	19.5	4.4

uments. First is the one that used transitive translation to measure the document word similarity score, shown in the first two rows. Second is the one that used the correct translation, the Japanese keywords contained in the Japanese queries, shown in the last two rows. This comparison shows that for the answer finder method, the use of transitive translation does not give significant effect which is different with the document retrieval results where the recall score of the transitive translation is about half of the one using the correct Japanese keywords.

As the final experiment, we applied the same answer finder module on our document retrieval results. We used the document retrieval with the combination of MI and TF $\times$ IDF filtering method. The question accuracy scores are shown in **Table 4**. We ranked the answers using the recall score (R) of the document retrieval and the text chunking (T) score resulted by Yamcha.

Even though our CLQA result is not well enough but it is higher compared to a similar research by QATRO[9] for English-Japanese CLQA with 0.5% accuracy for the Top1 answer. As comparison, we also conducted experiments on Indonesian-English CLQA using a similar technique as our Indonesian-Japanese CLQA. In the Indonesian-English CLQA, the final question answering accuracy achieves 22.5% while the passage retrieval recall score is 73.2%. Even though the English corpus used in the Indonesian-English CLQA is comparable with the Japanese corpus used in the Indonesian-Japanese CLQA, but the corpus size is different, the English corpus size is 17,741 articles and the Japanese corpus size is 658,719 articles. Other than the different characteristics between the English sentence and Japanese sentence (without word segmentation), this corpus size makes the Japanese document retrieval is more difficult than the English one. The main cause of low document retrieval score is the translation errors between Indonesian and Japanese. Using the available resources, the translation could not resolve the OOV proper noun problem. There are many proper nouns which are important question keywords could not be translated by our translation resources.

4. Conclusions

We have conducted an Indonesian-Japanese CLQA using easily adapted modules. By using transitive translation, the document retrieval results and the answer finder results are comparable with the translation of direct translation. In addition, compared to other research of English-Japanese using a similar approach, our Indonesian-Japanese CLQA gives better accuracy score. Even though the accuracy score is much worse than the Indonesian-English CLQA using the similar approach, but we believe that this result can be increased by improving the proper name translation. For the next research plan, we will try to improve the proper name translation method, for example by comparing the query keywords with a list of single kanji character.

Acknowledgments The work described in this paper was funded by a grant for "Frontiers of Intelligent Sensing" from the Ministry of Education, Culture, Sports, Science and Technology.

References

1) Adriani, M.: Using Statistical Term Similarity for Sense Disambiguation Cross-Language Information Retrieval, *Information Retrieval*, Vol.2, No.1, pp.71–82 (2000).
2) Ballesteros, L. A.: *Cross-language Retrieval via Transitive Translation*, Advances in Information Retrieval, chapter 8, pp. 203–230,

Kluwer Academic Publisher, 1st edition (2000).
3) Ballesteros, L.A. and Croft, B.W.: Resolving ambiguity for cross-language retrieval, *SIGIR 98: Proceedings of the 21st annual international ACM SIGIR conference on Research and Development in Information Retrieval*, New York, USA, ACM, ACM Press, pp.64–71 (1998).
4) Gao, J., Nie, J.-Y., Xun, E., Zhang, J., Zhou, M. and Huang, C.: Improving query translation for cross language information retrieval using statistical models, *SIGIR 01: Proceedings of the 24st annual international ACM SIGIR conference on Research and Development in Information Retrieval*, New York, USA, ACM, ACM Press, pp.96–104 (2001).
5) Gollins, T. and Sanderson, M.: Improving cross language retrieval with triangulated translation, *SIGIR 01: Proceedings of the 24st annual international ACM SIGIR conference on Research and Development in Information Retrieval*, New York, USA, ACM, ACM Press, pp.90–95 (2001).
6) Mori, T. and Kawagishi, M.: A Method of Cross Language Question-Answering Based on Machine Translation and Transliteration, *Proceedings of NTCIR-5 Workshop Meeting*, Tokyo, Japan, National Institute of Informatics, Nihon Printing Co., Ltd, pp.215–222 (2005).
7) Purwarianti, A., Tsuchiya, M. and Nakagawa, S.: Indonesian-Japanese Transitive Translation using English for CLIR, *Journal of Natural Language Processing, Information Processing Society of Japan*, Vol. 14, No. 2, pp. 95–123 (2007).
8) Qu, Y., Grefenstette, G. and Evans, D. A.: Resolving Translation Ambiguity using Monolingual Corpora, *Advances in Cross-Language Information Retrieval, Third Workshop of the Cross-Language Evaluation Forum CLEF 2002*, Rome, Italy, Cross Language Evaluation Forum, Springer, pp.223–241 (2002).
9) Sasaki, Y.: An Experimentally Extended QBTE Approach, *Proceedings of NTCIR-5 Workshop Meeting*, Tokyo, Japan, National Institute of Informatics, Nihon Printing Co., Ltd, pp.230–235 (2005).
10) Shimizu, K., Akiba, T., Fujii, A. and Itou, K.: Bi-directional Cross Language Question Answering using a Single Monolingual QA System, *Proceedings of NTCIR-5 Workshop Meeting*, Tokyo, Japan, National Institute of Informatics, Nihon Printing Co., Ltd, pp.236–241 (2005).