WHU Question Answering System at NTCIR-8 ACLIA Task

Han Ren
School of Computer,Wuhan University
129 Luoyu St., Wuhan, China
cslotus@mail.whu.edu.cn

Donghong Ji
School of Computer, Wuhan University
129 Luoyu St., Wuhan, China
donghong_ji@yahoo.com

Jing Wan
Center for Study of Language & Information, Wuhan University
129 Luoyu St., Wuhan, China
jennifer_wanj@yahoo.com

ABSTRACT
In this paper, we describe our system implemented for the NTCIR-8 CCLQA task. The system consists of a question translation model and a general question answering system for both factoid and complex questions. The translation model combines a translation engine and an online dictionary, which can provide more accurate translations of named entities in the questions. With regard to the question answering system, a PLSA based approach is introduced for answer sentence acquisition. For answer ranking, our system expands the question set by summarizing relevant sentences from a web knowledge base and leverages both semantic and statistical information of questions. In the official evaluation results, our system achieves 18.41% F-score in English to Chinese subtask and 25.66% in monolingual Chinese subtask.

Keywords
Cross-Lingual Question Answering, Complex Question, Document Retrieval, Probabilistic Latent Semantic Analysis, Answer Ranking

1. INTRODUCTION
The previous participation of WHU at CCLQA task focused on monolingual Chinese complex QA[7], in which patterns by manual work are adopted to acquire answer candidates. In NTCIR-8, complex questions are taxonomically extended as five types and factoid questions are also introduced, aiming at evaluating more effective QA systems. Since patterns are insufficient to satisfy the information requirements of questions with extended types, the system with less patterns or rules should be considered for a general performance.

Different with the previous system, we introduce a PLSA based answer sentence acquisition model to extract sentences that may contain the answers. By mapping sentences to a latent semantic space, sentences that semantically relevant with questions are selected and ranked as answer candidates. We also consider linguistic information in answer ranking and employ the web knowledge to expand the questions.

This year we participate in cross-lingual Chinese QA task. For this purpose, a translation model is implemented and EN-CS runs are submitted as well as CS-CS ones. In addition, we align our retrieval model with other IR4QA systems by submitting IR runs, aiming at evaluating the performance of our retrieval model. Although the retrieval documents are monolingual, our approach shows a general performance, which can deal with documents of any other language.

The rest of the paper is organized as follows. In Section 2 the system architecture is given. Section 3 gives the description of each model in detail. In Section 4, we give the experimental results and error analysis. Finally, the conclusion and future work are shown in Section 5.

2. System Architecture
The processing mechanism of our system is shown in Figure 1, which contains five main models: question translation, document retrieval, answer sentence acquisition, answer ranking and nugget extraction. The system carries out a pipeline approach with the following processes:

1) the questions are matched with the patterns to identify their types;

2) the key terms in the questions are extracted and translated, and then are submitted to the document retrieval model;

3) for each sentence in retrieval results, the answer sentence acquisition model is utilized to compute the similarity with the question; here we expand the question set by summarizing relevant sentences from a web knowledge;

4) after that, answer candidates are ranked by weighting their semantic and statistical similarity;

5) finally, the top 30 answers are selected to extract nuggets.

Details of each model are described in the following section.

3. Cross-Lingual QA System
In this section, we describe the pipeline approach, including question analysis, document retrieval and answer extraction and ranking, in our cross-lingual system. The monolingual system is almost same with the cross-lingual one except for the question analysis model.

3.1 Question Analysis
The goal of the question analysis is to determine the question type and to find key terms of a question. Translation model is also introduced for E-C task.

3.1.1 Question Type Identification
In NTCIR-8, not only complex question types are extended, but also factoid questions are added in the question set. Even we are

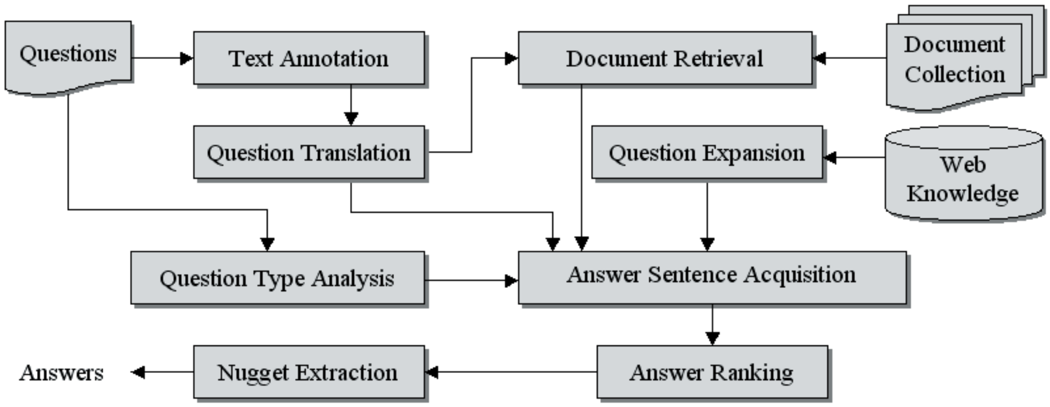

Figure 1. QA system processing mechanism

meant to follow the matching method introduced in our previous system, there are still some patterns that could not be utilized for identification because of multi-matching. For instance, in our previous system, the pattern that a question starts with '*Who be*' exclusively matches as 'BIOGRAPHY' type, whereas it could match another type of questions provided this year, i.e., the factoid questions asking for persons' names. We refine the pattern list and modify some heuristic rules so that the matching method can fit for the task of this year. Table 1 shows the pattern list. Note that the questions should be firstly dealt with POS tagging and Named Entity Recognition before matching.

Table 1. Question type pattern list

type	pattern
DEFINITION	什么是/是什么/**What is**
BIOGRAPHY	谁是+NE_{PER}/是谁+NE_{PER}/ **Who is+NE_{PER}**
RELATIONSHIP	和+的关系是什么 **the relationship/ interrelationship between**
EVENT	列举/举出/列出/说出/哪些
WHY	为什么 **Why/What is the reason**
PERSON	谁+vp-NE **Who+vp-NE**
ORGANIZATION	哪个/些+机构/部门 **Which**
LOCATION	哪里/哪个地方/哪些地方 **Where/Which+place**
DATE	何时/什么时候/哪一年 **When/What time/Which year**

Even the patterns can match most of questions accurately, there are still some multi-matching questions. We give some heuristic rules to get over the situation:

1) a longer pattern has a higher use priority than a shorter one if there is more than one pattern that matches a question;

2) if a question is matched by a pattern, it should not be matched with other patterns.

3.1.2 Key Term Extraction

For Chinese questions, we utilize ICTCLAS[1], a free Chinese lexical analysis tool that contains Word Segmentation, POS tagging and NER(Named Entity Recognition), to annotate them. Then all words and phrases except interrogation words and functional words are selected as key terms.

For English questions, the LBJ Tools [2], a free English linguistic analysis tool that contains POS Tagging, NE Chunking and NER, is utilized to annotate them. Although each model in it is independent, the results of POS Tagging have general format so that we can easily use them in the other two models. We also use a standard Porter algorithm for word stemming. Finally, NE chunks, named entities and other stemmed words except interrogation and functional ones are extracted as key terms.

3.1.3 Key Term Translation

We are not meant to translate the full text of each question, because our retrieval model treats each query as a bag of words, not considering the context of each word or the syntactic structure of each question. Initially, we employ some translation engines, i.e., Google Translate[3] and Yahoo Fanyi[4], but there are still some key terms that can not be translated correctly. For instance, the name of a movie is 'Riding Alone for Thousands of Miles' (ACLIA2-CS-0002) which means '千里走单骑', whereas the translation result is '单骑千里走' in Google Translate and '单独乘坐数以万计的英里' in Yahoo Fanyi. Some named entities are also mistranslated such as '陈忠雄' as the result of 'Sean Chen'. Alternately, we employ an online dictionary, ICIBA[5], to translate chunks and named entities. Although it is also based on statistical translation, the dictionary can give correct results of most chunks and named entities. Experimental results in the following section will show that the cross-lingual retrieval results that generated by the model that employs the online dictionary is close to the

[1] http://ictclas.org

[2] http://l2r.cs.uiuc.edu/~cogcomp/

[3] http://translate.google.com/

[4] http://fanyi.cn.yahoo.com/

[5] http://fy.iciba.com/

monolingual retrieval results. For other key terms, we still use Google Translate to deal with them.

3.2 Document Retrieval

Passage retrieval seems to be identified with most systems in CCLQA task, we still consider documents as the retrieval units. Our reason lies in two folds: 1) most documents are composed of short passages; some passages have only one sentence, and some passages even have several words, not a sentence. These passages may not be considered in passage retrieval in most cases or their ranks are very low, whereas some of them are potentially relevant to the questions. 2) our PLSA based answer sentence acquisition model described in the following section need enough semantically potentially relevant words to build the model, whereas passage retrieval could filter many of them. Based on the considerations above, we utilize documents as our retrieval results.

Table 2. Experimental results in NTCIR-7 dataset

method	M-AP	M-Q	M-nDCG
bigram+unigram+ BM25	**0.6061**	**0.6192**	**0.7956**
bigram+unigram+ VSM	0.5016	0.4873	0.6439
unigram+VSM	0.3897	0.3726	0.5292

Our retrieval model is based on Lucene, a free retrieval framework, and the index units are unigram and bigram. The original retrieval model in Lucene is VSM, whereas we find that the BM25 model that we employed in our system outperforms it in document retrieval. Table 2 shows the results of document retrieval in NTCIR-7 dataset. We can see from the table that BM25 model integrated with unigram and bigram promote the performance substantially.

3.3 Answer Sentence Acquisition

Pattern matching method is utilized in our pervious system for answer sentence acquisition. However, some patterns are out of date for the test set, and new patterns are more difficult to fetch since five more types of questions are added in this year. In addition, for a general performance, the answer sentence acquisition for factoid and complex questions should not be divided into two parts. Hence an acquisition model without hand-crafted rules should be considered, and it should have the ability to perform various types of questions.

For the complex questions, the potential answers mostly indicate the complex semantic relations with them so that the bag-of-word models are insufficient to compute the similarity between the answers and the questions. To solve it, some systems[1][6] employ LSA based models to build a semantic layer; thus the similarity between documents and words, or documents and documents are able to compute by using the semantically latent relations. However, the meaning of the decomposition algorithm is indefinite so that the model could lead to an uncontrolled performance for retrieval. For a clear decomposition meaning, PLSA(Probabilistic Latent Semantic Analysis) model[4] is introduced to build a semantic space of underlying topics, in which words and documents can be mapped as vectors. Some QA systems utilize PLSA to improve answer validation, i.e., modeling languages for document relevance estimation[2][3]. In our system, we employ PLSA model as the answer sentence retrieval model. Following is the description of the PLSA model in our system.

Given a sentence set *S*, a term set *W* and a topic set *Z*, the conditional probability of sentence-term *P*(*s*, *w*) can be described as follows:

$$P(s,w)=P(s)\sum_{z\in Z}P(w\mid z)P(z\mid s) \tag{1}$$

In (1), $P(w \mid z)$ represents the conditional probability of words in latent semantic layers(or topics), $P(z \mid s)$ represents the conditional probability of topics in sentences. Here the count for topic set *Z* is between 20 and 100. Then the model fits with the EM algorithm and export the optimal *P*(*Z*), $P(W \mid Z)$ and $P(Z \mid S)$. When a new query is coming, it is projected to the topic space *Z* by using EM algorithm. The similarity of the query and each sentence can be acquired by computing the similarity of the probabilistic distribution between them in the topic space.

Our algorithm is described as follows. First, initialize the *P*(*s*, *w*) for each sentence $s \in \{s_1, s_2, ..., s_n\}$ in the retrieved documents by using the ratio of the frequency of *w* in *s* and in the sentence set; and $P(w \mid z)$ for each word and $P(z \mid s)$ for each sentence are iteratively computed by EM algorithm. Then the query built from the question is mapped into the topic space to compute $P(z \mid q)$ by using EM algorithm, keeping $P(w \mid z)$ invariably. After that, the conditional probability $P(w \mid q)$ and $P(w \mid s)$ for each sentence in the retrieved documents is computed according to the formula (2):

$$P(w\mid s)=\sum_{z\in Z}P(w\mid z)P(z\mid s) \tag{2}$$

Finally, a *cosine* similarity method is utilized to compute the similarity between the query *q* and each sentence in retrieved documents, and the sentences that the weighting values are above a threshold are selected as the answer sentences.

$$Sim(s_1,s_2)=\frac{\sum_{w\in W}P(w\mid s_1)P(w\mid s_2)}{\sqrt{\sum_{w\in W}P(w\mid s_1)^2}\sqrt{\sum_{w\in W}P(w\mid s_2)^2}} \tag{3}$$

For a better performance, we submit each question to Wikipedia and summarizing the results as extended query sentences. The summarization approach is described in [9], which is suitable for summarizing in Wikipedia. Then we compute the similarity between each sentence in retrieved documents and query sentence set. Sentences that their similarities are above the threshold are selected into the answer set.

3.4 Answer Ranking

For a better performance of answer validation and ranking, some approaches combine various resources of evidence, but most of the resources are language-dependent and may not fit for complex questions. For a general performance, our method of answer ranking considers two main features: shallow semantic and statistical information of sentences. The motivation is, a sentence can be a potential answer candidate if 1) the sentence and the question have a semantically similar relation; and 2) most of the words in the sentence are also appear in questions or sentences derived from web knowledge bases. More specifically, although the retrieved sentences have the latent semantic similarity with questions, we should still consider the long distance dependency relations that could probably result in a low score for an answer candidate. On the other hand, a statistical similarity that treats the sentences as a bag of words could probably balance the impact of the semantic bias.

For semantic similarity, we consider the similarity of the main semantic roles that primarily profile the features of sentences. We choose the verb based labeling architecture derived from PropBank, in which 'predicate', 'subject', 'object' and some modifiers are the core roles for a sentence. For each sentence in answer candidate set we only label PRED, A0-A4, AM-LOC and AM-TMP and combine each predicate and the corresponding argument to a pair. The structure of a pair is described as follows:

$$\{w_1 \mid \text{PRED}, w_2 \mid \text{AM}\}$$

If a pair lies in both a sentence and a question, it is viewed as a matched pair. More specifically, every term and its semantic dependency relation should be matched if the pair is matched. Following is the weighting formula that we compute the semantic similarity:

$$M(s,q_i) = \frac{\#\ of\ matched\ pairs}{max\{\#\ of\ pairs\ in\ s,\ \#\ of\ pairs\ in\ q_i\}} \qquad (4)$$

For labeling of semantic roles, we utilize a system that we proposed in [8] to extract pairs. The system handles syntactic dependency parsing with a transition-based approach and utilizes MaltParser[6] as the base model. The system also utilizes a Maximum Entropy model to identify predicate senses and classifies arguments.

For statistical similarity, we simply utilize a cosine similarity to compute it. The weighting formula of our method for answering ranking is as follows:

$$w_s = max\{\alpha \bullet M(s,q_i) + (1-\alpha) \bullet CosSim(s,q_i)\} \qquad (5)$$

α is an adjusting parameter, and q denotes each sentence in the query sentence set described in section 3.3. According to (5), the sentence that mostly similar to a query is acquired. Finally, 30 answer sentences is extracted and ranked by their ranking scores.

3.5 Nugget Extraction

Our nugget extraction model is simple. For factoid questions, we extract named entities according to the type of them. For instance, if a question is classified as the type of 'PERSON', we extract the person names from the ranked answer sentences and rank them with the scores of the sentences. For complex questions, we preserve the sentences as the answers.

4. Experimental Results and Analysis

We submit three types of runs in this year's task: question analysis, IR4QA and CCLQA. Because of time limitation, we do not participate in the IR4QA+CCLQA subtask.

4.1 Question Analysis Runs

We submit three runs for this subtask. E-C-01-T and C-C-02-T utilize the exact methods given in section 3.1. E-C-03-T is very close to E-C-01-T except that multiple translation results using the online dictionary is added as the key terms. The motivation is that since a term probably has the multiple translation results, they should be reserved for retrieving more relevant documents, such as the name 'Bin Laden'(ACLIA2-CS-0006), which can be translated as '本拉登' or '宾拉登'. IR4QA results indicate that E-C-03-T achieves a better performance than E-C-01-T. We also find that the latter translation result of the name seldom appears at the documents. Hence the frequency should be considered when selecting multiple translation results of a word. In addition, some words are not translated correctly, such as 'screen'(ACLIA2-CS-0002), which is mistranslated as '甄别', and that's why the retrieval results using E-C translation runs achieve a lower performance than using C-C runs. It is also notable that the performance of question type identification in E-C run outperforms that of in C-C run. For instance, the question ACLIA2-CS-0006 in English is easy to identify because it matches a pattern of the type 'RELATIONSHIP' in Table 1, whereas the corresponding Chinese question can not matches any pattern of this type. It mainly because that patterns in some types are inaccurate. Moreover, some patterns probably lead to over-matching. But most of them are still appropriate in our system. Hence the patterns should be refined for a better identification performance.

4.2 IR4QA Runs

Retrieval model may impact the performance of the QA system. To investigate the appropriate IR strategy that brings out the superior QA performance, CCLQA participants are also required to submit their retrieval results with IR4QA participants. Due to 27 invalid topics in which the number of relevant documents found in the depth-100 pool was fewer than five, only 73 topics are utilized in the evaluation set.

In this year, we submit 2 runs for E-C and C-C task respectively. For E-C runs, we investigate the impact of two translation strategies to retrieval results. Run 1 utilizes both the translation engine and the online dictionary mentioned in section 3.1.3. Run 2 only utilizes the translation engine. For C-C runs, different retrieval units and models that impact the results are checked. Run 1 utilizes both unigram and bigram as the retrieval units, and run 2 just utilizes bigram. The evaluation results for all runs are shown in Table 3.

Table 3. The official IR4QA results of our system(before bug fix)

	M-AP	rank	M-Q	rank	M-nDCG	rank
E-C-01-T	0.371	15	0.4085	15	0.6139	14
E-C-02-T	0.3555	16	0.393	16	0.5989	17
C-C-01-T	**0.4128**	5	**0.4528**	4	**0.6629**	4
C-C-02-T	0.4055	8	0.4458	8	0.6577	6

Ranks in Table 3 are based on all runs for IR4QA task. Actually, in E-C subtask, our best run achieves the rank 5 of M-AP, 5 of M-Q and 4 of M-nDCG; in C-C subtask, our best run achieves the rank 4 of M-AP, 4 of M-Q and 4 of M-nDCG. The ranks prove that our retrieval model is effective to some extent. From the offi-

[6] http://w3.msi.vxu.se/~jha/maltparser/

Table 4. The official CCLQA results of our system

EN-CS	**Recall**	**Precision**	**F-score**
Definition	0.1967	0.2228	0.1033
Biography	0.5238	0.0678	0.2967
Relationship	0.1917	0.0503	0.1010
Event	0.3113	0.0401	0.1557
Why	0.0671	0.0072	0.0218
Person	0.8000	0.1200	0.4435
Organization	0.6000	0.3211	0.5437
Location	0.8000	0.3780	0.4849
Date	0.4000	0.1601	0.2952
Overall	0.3161	0.0775	0.1841

CS-CS	**Recall**	**Precision**	**F-score**
Definition	0.3802	0.0380	0.1885
Biography	0.9167	0.1223	0.5313
Relationship	0.1583	0.0924	0.1239
Event	0.3855	0.0367	0.1807
Why	0.1077	0.0092	0.0418
Person	1.0000	0.2632	0.7731
Organization	0.6000	0.2978	0.5407
Location	0.6000	0.3501	0.4145
Date	0.8000	0.1893	0.5795
Overall	0.4100	0.0987	0.2566

cial IR4QA results, we also find some situations that need to be discussed.

- The improvement of performance(1.55% of M-AP, 1.55% of M-Q, 1.5% of M-nDCG) from E-C-02-T to E-C-01-T is derived from the utilization of the online dictionary mentioned in section 3.1.3. On the other hand, although the translation engines mistranslate some terms, the retrieval model still return most correct documents that include the answer sentences. For instance, we utilize the different translation results of ‘Riding Alone for Thousands of Miles’ by translation engines and online dictionaries to retrieve documents, and most of documents, especially correct ones within two retrieval results, are matched. In other words, the word order in long terms does not play a very important role if the key words are correctly translated. Therefore, the main effect of online dictionaries is to translate some key terms or named entities, i.e., the person names and the location names.
- As to C-C subtask, we investigate the impact of different retrieval units and parameters in BM25 model to the results. It is notable that the impact for the performance is slight(0.73% of M-AP, 0.7% of M-Q, 0.52% of M-nDCG) since C-C-01-T adopts unigram and bigram, and C-C-02-T only utilizes bigram as the retrieval units. It implicates that only using bigram can achieves a good performance.
- In comparison with other participants that submit both the E-C and the C-C results, we find that the positive divergences of the performances between our E-C and C-C result is smallest[7]. It indicates that our translation strategy outperforms the other participants in some extent.

4.3 CCLQA Rums

Since the human evaluation of CCLQA considers the first priority runs submitted by participants, we only submit two runs for this task, based on the RUN-1(tagged as 01-T) of question analysis and IR4QA. Table 4 shows the official results for each type of questions by Recall, Precision and F(β=3) score.

For C-C task, the F-score over all questions is 25.66%, ranking second in all submitted runs[5]. For E-C task, we rank first as the F-score 18.41%, because no runs are submitted for evaluation except us.

From the official CCLQA results, some conclusions can be drawn:

- The F-scores of the factoid questions greatly outperforms that of the complex questions in both E-C and C-C task. Although the factoid and complex questions are not performed respectively in the retrieval and answer acquisition model, our system is still very available for factoid ones. It indicates that these two models in our system have a general performance for these two kinds of questions.
- Our system achieves a very low performance for the question of the type ‘Why’ in both E-C and C-C task. It is mainly because that the semantic relations in these questions are more complex than others. Actually, most answers have the logic relations with the question, rather than synonyms or shallow semantic ones. For instance, when asking ‘Why is "ShenZhou" spacecraft launched in relatively cold season’, the answer is most like a reason chain: the spacecraft is recovered by the survey vessels; the survey vessels are mainly located in the southern hemisphere; the recovery task befits in summer; the climate is quite opposite between the southern and the northern hemisphere. However, most words in the answer do not appear at the question. Alternately, sentences in answers have logical(or casual) relations that should be inferred from one to another. The answer to the question can be acquired only if all the relative events are chained through inference. Hence the inference models should be considered for a better performance of ‘Why’ questions.
- As a close view to the results we find that, in most cases the Recall scores greatly outperform the Precision ones for each type of the questions. And the reason is easy to know: the nuggets extracted are too long in comparison with the human results. In addition, the scores of the factoid questions greatly outperform that of the complex ones, and this is mainly because the short answers(mostly are named entities) are extracted as the nuggets for the factoid questions whereas most answer sentences are remained as the nuggets for complex questions. Hence if the system is required for a better performance, the nugget extraction

[7] We consider the divergence of the performance between E-C and C-C results by using a score, i.e., M-AP, in C-C results to minus the corresponding score in E-C results. If the score is greater than zero, it is named positive.

model should be improved to acquire more fine-gained answers.

5. Conclusion

In this paper, we describe our system implemented for NTCIR-8 CCLQA task. The main contribution to the system lies in two folds: 1) a PLSA model based answer acquisition approach is introduced for both factoid and complex questions; and 2) semantic and statistical information are combined in the answer ranking model for a better performance. Although a general performance is achieved according to the official evaluation results, we still face two important problems: 1) with the improvement of the complexity of questions, deep knowledge in them should be acquired by some inference mechanisms; 2) more refined nugget extraction model are needed to improve the precision of the system.

6. ACKNOWLEDGMENTS

This work is supported by Natural Science Foundation of China(Grant Nos. 60773011, 90820005), Wuhan University 985 Project: Language Technology and Contemporary Social Development(Grant No.985yk004) and Self-research program for Doctoral Candidates of Wuhan University in 2008.

7. REFERENCES

[1] Yu Bai, Li Guo, Dongfeng Cai and Bo Zhou. 2008. KECIR Question Answering System at NTCIR7 CCLQA. In proceedings of the 7th NTCIR Workshop meeting, Tokyo, Japan, 54-59.

[2] Protima Banerjee and Hyoil Han. 2008. Incorporation of Corpus-Specific Semantic Information into Question Answering Context. In proceedings of CIKM 2008 Ontologies and Information Systems for the Semantic Web Workshop, Napa Valley, CA.

[3] Protima Banerjee and Hyoil Han. 2009. Answer Credibility: A Language Modeling Approach to Answer Validation. In proceedings of NAACL-HLT 2009, Boulder, Colorado, 157-160.

[4] T Hofmann. Probabilistic latent semantic indexing[C]. SIGIR-99, Berkeley, CA, 1999.

[5] Teruko Mitamura, Hideki Shima, Tetsuya Sakai, Noriko Kando, Tatsunori Mori, Kohichi Takeda, Chin-Yew Lin, Ruihua Song, Chuan-Jie Lin and Cheng-Wei Lee. 2010. Overview of the NTCIR-8 ACLIA Tasks: Advanced Cross-Lingual Information Access. In proceedings of the NTCIR-8 Workshop Meeting. Tokyo, Japan.

[6] Tetsu Muramatsu and Tatsunori Mori. 2004. Integration of PLSA into Probabilistic CLIR Model. In proceedings of the 4th NTCIR Workshop meeting, Tokyo, Japan.

[7] Han Ren, Donghong Ji, Yanxiang He, Chong Teng and Jing Wan. 2008. Multi-Strategy Question Answering System for NTCIR-7 C-C Task. In: Proceedings of the 7th NTCIR workshop meeting, Tokyo, Japan, 49-53.

[8] Han Ren, Donghong Ji, Jing Wan and Mingyao Zhang. 2009. Parsing Syntactic and Semantic Dependencies for Multiple Languages with A Pipeline Approach. In proceedings of the Thirteenth Conference on Computational Natural Language Learning, Boulder, Colorado.

[9] Shiren Ye, Tat-Seng Chua and Jie Lu. 2009. Summarizing Definition from Wikipedia. In proceedings of the 47th Annual Meeting of the Association for Computational Linguistics. Singapore.