KECIR Question Answering System at NTCIR7 CCLQA

Yu BAI，Li GUO，Lei LIU，Dong-feng CAI，Bo ZHOU
Knowledge Engineering Research Center, Shenyang Institute of Aeronautical Engineering
Email: nlpxiaobai@yahoo.com

Abstract

At the NTCIR-7 CCLQA (Complex Cross-Language Question Answering) task, we participated in the Chinese-Chinese (C-C) and English-Chinese (E-C) QA (Question Answering) subtasks. In this paper, we describe our QA system, which includes modules for question analysis, document retrieval, information extraction and answer generation. Besides, we used an online MT (Machine Translation) system to deal with question translation in our E-C task. An overall analysis and a detailed module-by-module analysis are presented. Since document retrieval is an essential part of CLQA, we also did experiments and submit QA results using IR4QA results in order to find out which IR technique would help CCLQA.

Keywords: *Complex Question answering, Question analysis, Answer generation*

1. Introduction

According to an advance in Natural Language Processing technology, Question Answering has become a popular research field in computational linguistics [1].

Current research in QA is moving beyond factoid questions, so there is a significant motivation to evaluate more complex questions in order to move the research forward [2]. We participated in the Chinese-Chinese (C-C) and English-Chinese (E-C) QA (Question Answering) subtasks at the NTCIR7, the task evaluates research on four types of questions: events, biographies, definitions, and relationships.

We follow a modular architectural approach to QA depicted in Figure.1: For a given question, we did question classification first, the question was classified into its expected answer category by using pattern matching with predefined templates, and then, a query form was generated by the original question, it could be used to retrieve relevant documents from the target collection. A few sentences were extracted from those relevant documents to form a candidate pool. Here, a scoring approach was used to rank candidate answers in this period. For English-Chinese CLQA, an online MT (Machine Translation) system was employed to render the original English question into a Chinese query.

Since document retrieval is an essential part of CLQA, we also did experiments and submit QA results using IR4QA results provided by the teams participated in IR4QA task in order to find out which IR technique would help CCLQA.

The remainder of this paper is organized as follows. Section 2 describes our Chinese (Simplified) monolingual QA methods and results, and Section 3 describes our E-C experiments and comparison with monolingual results. Section 4 has some additional experiments. Finally, Section 5 has our conclusions.

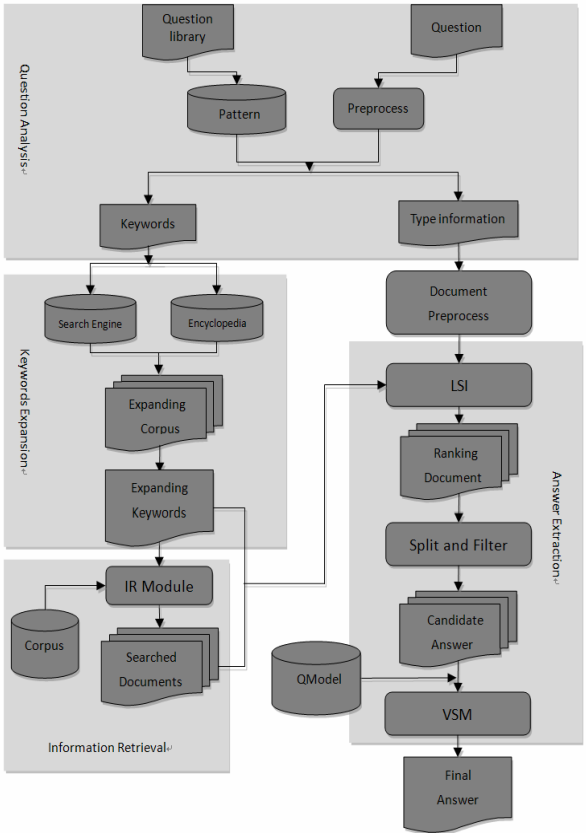

Figure1. C-C QA architectural module

2. Chinese Monolingual QA

The aim of cross-lingual information access is to answer any type of question or to retrieve any type of information needs in any language with responses drawn from multilingual corpora. If the information needs are very simple ones (e.g. factoid question), then the answer can be a simple word or phrases. If the information needs are more complex, then the answers may come from multiple documents [3].

In order to find more distinct answers of the complex question, we employ a 4-dimensional QA Event Space [4] which consists of Human/Object, Time, Location and Action axes. Take the question "*What is the impact of the change of the Oil prices?*" as an example, it can be

projected to Time, Location, Action and Object four axes, as illustrated in Figure2. And the process of answering complex questions can be regarded as calculation of coordinates in Complex Question Model.

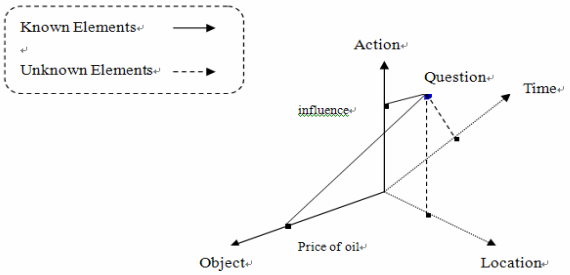

Figure2. Complex Question Model

As shown in Figure1, The work flow of our QA system is as follows:

(1). Question analysis. The question and answer type are acquired through pattern matching approach, while the keywords of question are projected in the Event Space.

(2). Question expansion. After question analysis, we got some basic query keywords, and then to get more reasonable answers, a web based keywords extension approach has been used, a series of query forms could be generated in this step.

(3). Document retrieval. This module use IR query generated by those above steps as inputs, and retrieve relevant documents from the target collection.

(4). Answer extraction. In order to extract the final answer(s), we adopt the method which combined Latent Semantic Indexing with Complex question model to calculate similarity between question and candidate answers.

2.1. Question analysis

This module is used to determine the question type, the answer type and the keywords by using question pattern. The complex questions can be classified to Event, Relationship, Biography and Definition in accordance with the question type, whereas the answer type is the main interrogative intention of the question.

2.1.1. Question pattern library. To build our complex question corpus, we collected complex questions from the Web. Then, we use the named entity type to replace the entity identified by NE recognition. For example, for the question:

张朝阳和搜狐有什么关系

(What's the relationship between Charles Zhang and sohu)

We use <PERSON> to replace Name *Charles Zhang* and use <FIRM> to replace *sohu*. Then we got the pattern :

<PERSON>和<FIRM>有什么关系

(What is the relationship between <PERSON> and <FIRM>)

A semantic can be expressed in a variety of ways, especially in Chinese. Therefore, we collected the questions as many as possible for these expressions. Question-equivalent expression of every Q-type is built manually in order to produce more patterns (Table 1).

However, many questions can't be matched because of wildcard queries, e.g. the question "*Which metropolis are <RIVER> flowed through?*" or "*Which provinces are <RIVER> flowed through?*" can't match the pattern "*Which cities are <RIVER> flowed through?*". "metropolis" and "cities" are synonyms, while "provinces" and "cities" are in same entity class, So we expanded the non-entity word in the pattern to constituent new pattern. Finally, we obtain the Chinese Complex question patterns (Table 2).

Table1. Question-equivalent expression

Q-type	equivalent expression
Eve	A 的 B 有哪些<==>A 有哪些 B<==>列举 A 的 B…
Bio	谁是 A<==>A 是谁…
Rel	A 和 B 是什么关系<==> A 和 B 有哪些关系…
Def	什么是 A<==> A 是什么…

Table2. Chinese Complex question patterns

Q-type	equivalent expression
Eve	请 $[列举/举出/罗列/历数/…] <EVENT> 对 <LOCATION> [的] $[影响/改变/危害/…][有哪些/是什么/有什么/是哪些]
Bio	请 $[介绍/告诉/简介…]; 谁是<PERSON>
Rel	请 $[介绍/告诉/简介…]<PERSON> 和 <PERSON> $[有/是][什么/哪些]关系
Def	请$[介绍/告诉/简介…]?什么是*

2.1.2. Question classification. The purpose of question classification is to obtain the question type and the answer type. The classification of a question can be achieved through the matching of question patterns. e.g, "*谁是邓肯?*" is a biography question and the answer type is <Object>, because of matching the pattern "*谁是* <PERSON>". But unavoidably, some questions don't match any patterns we collected. There, the question type and answer type are determined by some rules, such as the words which represent the obvious information of types, the named entities, e.g. "*举出为抗洪救灾捐款的个人。*" can not be matched with any of our question patterns. But it has "*举出*" which is the keyword of event question, and the answer type is determined by the last word of the question. "*个人*" suggests that the answer type of this question is Human.

2.1.3. Keyword extraction. If the question matches the pattern, the keywords are extracted by the positions where are in the pattern, e.g., in the pattern "*What is the relationship between <FIRM> and <PERSON>*". The keywords are the words represent <FIRM> and *<PERSON>* and "*relationship*". Otherwise, the keywords are the lift words after flitted stop words.

2.2. Keywords expansion

We can get some keywords in question analysis step, but they only appear in the question, and it's possible for them to appear in corpus in another way, especially in

Chinese. For instance, the keywords in *"谁是本拉登？"("Who is Osama bin Laden?")* are *"本拉登"*("*Osama bin Laden*"), but "*本-拉登*", "*拉丹*", "*拉登*", *"宾拉登"*, these maybe appeared in corpus all mean *"本拉登"*. Thereby we must find out the synonyms of the keywords as far as possible, so as to improve recall of our system. This module uses online encyclopaedia and search engine as our expanding source.

We used an online encyclopedia as the knowledge resource for the keywords expansion of Biography and Definition questions. For these words, another web search engine (http://www.g.cn) has been employed. We got the top 30 fragments returned by using those words and the documents expressed in online encyclopedia together to compose the expanding corpus. Then we use tf*idf score to extract the keywords from it. In order not to loose some important information, we extract the entities and the nouns which are near to the entities(less than 5 words) as supplementary. For example, we can get "*乌萨玛*" "*拉登*" "*宾拉登*" "*拉丹*" "*沙特*" "*美国*" "*9.11*" "*恐怖分子*"…from the passage which introduced Osama bin Laden(illustrated in Figure3).

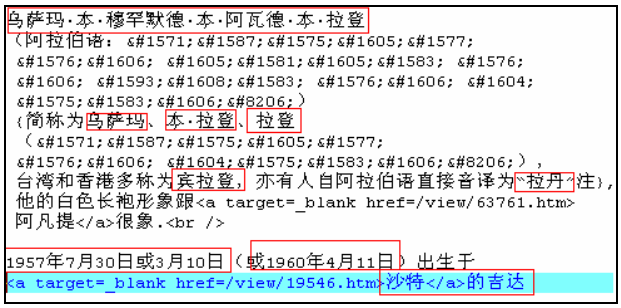

Figure3. Keywords Expansion example

These expanding words and original keywords together compose a new keywords set, we call it Question Terms.

2.3. Document retrieval

Since the ACLIA task also contains a cross-lingual information retrieval (CLIR) task, called IR4QA (Information Retrieval for Question Answering). During the evaluation, the question text and QA system question analysis results were provided as input to the IR4QA task, which produced retrieval results that were subsequently fed back into the end-to-end QA systems.

Our group also participated in the IR4QA task at NTCIR7, and it aim to evaluate traditional ranked retrieval of documents containing answers. For our QA System, we use this as the document retrieval module.

For details on our IR4QA system, we refer the reader to the KECIR IR4QA system description paper[5].

2.4. Answer extraction

This module focuses on extracting final answer from the document retrieved from results. The document ranking doesn't mean the answer ranking. Top document does not necessarily contain sentence that is an answer, and the lower ranking of the document may also contain the right answers, so not only the similarity between the question and every document(S_{QD}) is computed, but also the similarity between the question and every sentence(S_{QS}) in the document(here, question is question terms).

(1). the measure of S_{QD}: Latent Semantic Index(LSI) can map documents and question terms into a lower dimensional space composed of higher level concepts which are fewer in number than the keywords, besides it has a higher performance than Vector Space Retrieval which may contain noisy and vague information. So we use it to calculate S_{QD}. First, the question terms and the documents retrieved are transformed to a document-keywords matrix like Table 3 where the row vector represent the document, and the column vector represent the term which has appeared in both the documents and the question terms.

Table3. Document-Keywords Matrix
(QT: Question Terms D:document K: keyword)

D K	K_1	K_2	…	K_n
QT	X_{11}	X_{12}	…	X_{1n}
D_1	X_{21}	X_{22}	…	X_{2n}
…	…	…	…	…
D_m	X_{m1}	X_{m2}	…	X_{mn}

Then we apply log and entropy transformations to the matrix[6]. Let $X_{i,j}$ be the cell in row i and column j of the D-K matrix, which means the frequency of K_i appears in D_j. We wish to weight the cell $X_{i,j}$ by the entropy of the j-th column. To calculate the entropy of the column, we need to convert the column into a vector of probabilities. Let $p_{i,j}$ be the probability of $X_{i,j}$, detailed in Formula a. The entropy of the j-th column is then H_j, detailed in Formula b. For all i and all j, replace the original value $X_{i,j}$ by the new value $[1-H_j/\log(m)]\log(X_{i,j}+1)$. This is an instance of TF-IDF family of transformations, $\log(X_{i,j}+1)$ is the TF term and $[1-H_j/\log(m)]$ is the IDF term[7].

$$p_{i,j} = x_{i,j} / \sum_{k=1}^{m} x_{k,j} \qquad \textbf{(a)}$$

$$H_j = -\sum_{k=1}^{m} p_{k,j} \log p_{k,j} \qquad \textbf{(b)}$$

Finally, S_{QD} is calculated through cosine angle between the row vectors of the document and the question in the matrix processed by LSI.

(2). the measure of S_{QS}: After anaphora resolution, we segment every document to sentences by full stop and filter the stop words, and the sentences which contain no keywords in the question will be eliminated, thus the candidate answer sentences set is produced.

We calculate the similarity S_{QS} through Vector Space Model (VSM) in the complex question model, in order to find the sentence which contain comprehensive information. After projecting the candidate answers and the question terms to our complex question model, we calculate the cosine angle through VSM.

For example, there are three candidate answers to the question "*列举出与北京大学百年校庆相关的大事*" , one of them is "*北京大学这所中国的最高学府订于 5 月 4 日庆祝建校 100 周年，当天上午将在人民大会堂隆重集会，举*

行“百年校庆”庆祝大会”, the other is “*中国领导人江泽民参加了在人民大会堂举行的北大建校百年庆祝大会。*”, and the last is “*北京图书馆今天在京收藏“北京大学百年华诞纪念卡”第 1898 号卡和第 100 号卡*”.

The final answer is determined by the value $S_{QD} \times S_{QS}$, the greater value, the higher ranking.

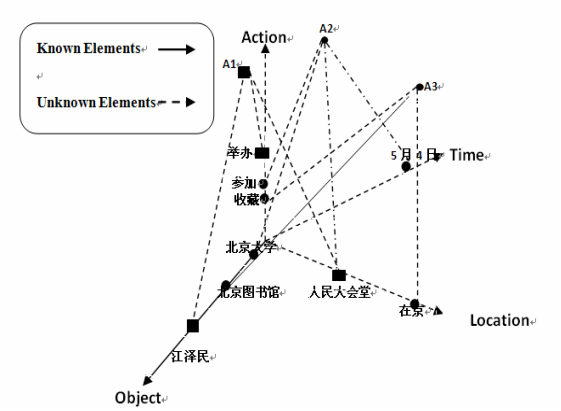

Figure4. Similarity in Complex Question Model

2.5. C-C results & analysis

Due to the problem of our submission, the overall submission to the Chinese-Chinese (C-C) subtask only included one run in NTCIR-7 CLQA task. But in the CCLQA using IR, we have submitted 20 C-C runs which contain three runs using our results of IR4QA. Detailed in Table 4.

Table4. C-C results

Run	Def	Bio	Rel	Eve	All
KECIR-CS-CS-01-T	0.241355	0.158935	0.223187	0.122077	0.183637
KECIR-CS-CS-02-T	0.241355	0.158935	0.223187	0.122077	0.183637
KECIR-CS-CS-03-DN	0.251265	0.17304	0.218307	0.120367	0.186463
CS-CS-02-DN	0.2213	0.1262	0.0143	0.0801	0.0978

The first three runs use the results of IR4QA, whose differences are described in IR4QA paper. And in the last run, we only use some answer patterns to extract the answer form the definition questions and VSM to calculate the similarity between the answer and the question for the other three kinds of questions. Interestingly, the method only using pattern has a better performance than our image. Since to the definition question, the number of the expressions of the answer is small. For example, the topic 378 “*什么是黑洞?*” which has six final answers, we can use the answer pattern “*keyword+[是/又叫/也称为/即] **” to exact four final answers directly, so pattern is good at processing the definition question.

Besides first three runs listed above, we have submitted 17 runs using the results of IR4QA (Table 5.) It can be seen that our system is good at definition question relatively. We consider that there are mainly two reasons:

(1).The effect of keywords expansion, especially using Chinese encyclopaedia as expanding resource, because the purpose of definition question is to find the explanation of the keyword, and what the encyclopaedia records is all most explanations of term, so keywords expansion plays more important role than the other types of questions.

(2).The effect of answer patterns, this is the biggest difference with the Biography question, because the latter is suitable for keywords expansion using encyclopaedia as expanding resource too.

Table5. IR4QA+CCLQA collaboration track (C-C)

Run	Def	Bio	Rel	Eve	All
CMUJAV-CS-CS-01-T	0.215755	0.18393	0.19857	0.118933	0.175188
CMUJAV-CS-CS-02-T	0.229155	0.177655	0.204257	0.14349	0.185686
NLPAI-CS-CS-01-T	0.357395	0.135305	0.185767	0.120767	0.1905
NLPAI-CS-CS-02-T	0.36704	0.109735	0.217333	0.10971	0.193468
NLPAI-CS-CS-03-T	0.36704	0.095195	0.197257	0.108927	0.184302
NLPAI-CS-CS-04-T	0.383205	0.101825	0.202363	0.0890933	0.184443
NLPAI-CS-CS-05-DN	0.36704	0.119675	0.216983	0.10485	0.193893
OT-CS-CS-01-T	0.23113	0.149865	0.194813	0.0977967	0.163982
OT-CS-CS-02-T	0.25092	0.19056	0.21183	0.103867	0.183005
OT-CS-CS-03-T	0.236245	0.18066	0.193127	0.08848	0.167863
OT-CS-CS-04-T	0.24889	0.1937	0.193887	0.10285	0.177539
OT-CS-CS-05-T	0.25515	0.184115	0.199473	0.10046	0.177833
RALI-CS-CS-01-T	0.176905	0.167335	0.198087	0.121717	0.164789
RALI-CS-CS-02-T	0.214725	0.180335	0.205733	0.135667	0.181432
RALI-CS-CS-05-T	0.22106	0.168455	0.208623	0.136337	0.181391
WHUCC-CS-CS-01-T	0.22337	0.181395	0.206	0.144583	0.186128
WHUCC-CS-CS-02-T	0.24344	0.185845	0.206907	0.141823	0.190476
Average	0.26991	0.159152	0.202412	0.115844	0.181289

3. English-Chinese CLQA

The process to deal with the questions in English is the same to the process to deal with the questions in Chinese, except question pattern acquirement and keyword translation.

3.1. Question pattern acquirement

English question pattern is acquired by the human translation of Chinese question pattern, for example, Chinese question pattern “*谁是**” can be translated to English question pattern “*Who is **”.

3.2. Keyword translation

We use online MT system to translate the keywords extracted from question analysis. For example, “gas hydrates ” is translated to “*天然气水合物*”. However, the translation of name is different from the others, A is a English name, and A is translated to B through the online MT system, then we translate B to C, if A=C, B is the final translation of A, if A≠C, we use A as a query in search engine, and the recommendation of

query given by search engine is the final translation of A. For example, "*Yang Liping*" is translated to "*杨哩瓶*", and "*杨哩瓶*" is translated to "*Yang miles bottle*" which is different from "*Yang Liping*", so we input "*Yang Liping*" as query, and the search engine return "*Do you want to find 杨丽萍?*", and the red characters are the final translation.

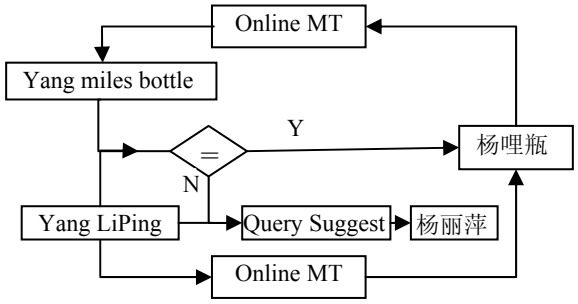

Figure5. Translation correcting

3.3. E-C results & analysis

our overall submission to the English-Chinese (EC) subtask included two runs in NTCIR-7 CLQA task(detailed in Table6). "EN-CS-01-T" uses the method which we have described above, "EN-CS-02-T" uses the stage without any of the three innovations, and in the CCLQA using IR, we have submitted 18 EN-CN runs(detailed in Table7).

Table6. E-C results (E-C)

Run	Def	Bio	Rel	Eve	All
EN-CS-01-T	0.2452	0.2563	0.1584	0.1364	0.1887
EN-CS-03-T	0.2154	0.2752	0.0692	0.0732	0.1408

And we can see that the our system has a better performance in Definition and Biography questions relatively. Besides the effect of answer pattern and keywords expansion, the most important reason is the translation of keywords, since most of the keywords in Definition, Biography and Relationship questions are entities, such as names and short phrases which only include two or three words. For example, "*Bill Gates*"(*比尔盖茨*), "*Big Bang Theory*"(*宇宙大爆炸理论*), "*black hole*"(*黑洞*) and so on. However, the keywords in Event questions are phrases which contain some structural information. For example, "*the US government suing Microsoft for its anti-competitive business behaviors*"(*美国政府控告微软反垄断商业行为*), "*the birth of cloned animals in the world*"(*世界上的克隆动物*) and so on. The translation is described in Figure5.

4. Post-Evaluation Experiments

In order to show the performance of our system, we do two experiments as supplement.

4.1. Keywords translation experiment

In the subtask of E-C CLQA, MT is very important to the result. Especially, the translation of keywords are fatal to performance. So we show our translation of keywords in table , and the "Complete Match" means the translation of English keyword is the same as the keywords in Chinese question literally, and the "Human evaluate" means the translation of English keyword express the same meaning as the keywords in Chinese question semantically.

Table7. CCLQA using IR (E-C)

Run	Def	Bio	Rel	Eve	All
CMUJAV-EN-CS-01-T	0.232455	0.167255	0.162747	0.113937	0.162947
CMUJAV-EN-CS-02-T	0.22922	0.17399	0.17917	0.11352	0.168449
CYUT-EN-CS-01-T	0.219325	0.14711	0.14039	0.111967	0.148994
CYUT-EN-CS-02-D	0.242415	0.166645	0.15623	0.12132	0.165077
CYUT-EN-CS-03-DN	0.240635	0.12094	0.166837	0.104973	0.153858
HIT-EN-CS-01-DN	0.201825	0.16718	0.167137	0.124947	0.161426
HIT-EN-CS-02-D	0.24091	0.160965	0.19691	0.140613	0.181632
HIT-EN-CS-02-DN	0.201825	0.16718	0.167137	0.124947	0.161426
MITEL-EN-CS-01-T	0.19906	0.167765	0.1799	0.1177	0.162645
MITEL-EN-CS-02-T	0.240515	0.155715	0.19816	0.113743	0.172817
MITEL-EN-CS-03-T	0.25733	0.16108	0.21014	0.120633	0.182914
MITEL-EN-CS-04-D	0.240815	0.163295	0.19948	0.109573	0.173538
MITEL-EN-CS-05-TD	0.240815	0.169695	0.198933	0.107957	0.174169
MITEL-EN-CS-06-T	0.2288	0.195795	0.18653	0.109987	0.173874
RALI-EN-CS-01-T	0.17627	0.13939	0.17689	0.108457	0.148736
RALI-EN-CS-02-T	0.22302	0.153125	0.183127	0.109867	0.163127
RALI-EN-CS-04-T	0.235345	0.150395	0.174127	0.111313	0.16278
RALI-EN-CS-05-T	0.218515	0.141665	0.17396	0.0966233	0.153211
Average	0.226060	0.159399	0.178766	0.11455985	0.16509

Table8. Keywords Translation

	Def	Bio	Rel	Eve
Complete Match	85%	85%	80%	56.7%
Human	95%	100%	86.7%	70%

It can be seen that the translation of keywords in definition, biography, relationship questions are better , and it is one of the reasons why these questions performs better than the event ones in E-C result(Table8).

4.2. Question classification experiment

In Question Analysis, the classification of question types is very important to us, and we use the pattern matching to finish it in our system. In order to produce more useful patterns, we adopt a method of pattern expansion, and this experiment show us the difference among the unexpanded, the expanded and the expanded + rule.

The value in the figure represent the number of sentences which are correctly classified.

As shown in Figure6, the number of matched questions using expanded patterns is more than using unexpanded patterns, so the expanded pattern has higher performance than the unexpanded ones.

Particularly, the number of expressions alters along with the classification of the question. To the definition and biography question, the unexpanded patterns are enough to match all the questions, because these questions are all "*什么是**" or "**是什么*" for definition question, and "**是谁*" or "*谁是**" for biography question; To the relationship question, the expressions are little more than the former question, e.g. topic83: "*举出通货膨胀与经济的关系。*", however, the unexpanded pattern which is the most similar to the question is "*[列举/列出] A 和 B 的关系*", and the expanded pattern "*举出 A 和 B 的关系*" can be produced through synonyms expansion; To the event questions, the number of the expressions is the most, so the number of the pattern which is matched is the least. e.g. topic82 "*举出悉尼奥运会创了哪些新高。*" isn't able to match any of the patterns, and the question type is determined by the keyword "*举出*".

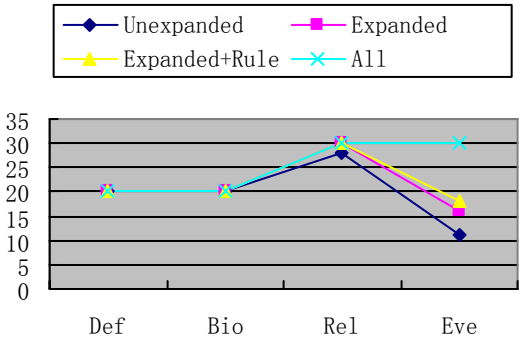

Figure6. Question Classification Results

5. Conclusion

We have described our CCLQA system at the Nticr-7. And we have evaluated the result of the official run and post-evaluation experiments. Our method is effective, but there are still some problems: 1. Keywords expansion is fatal to our system, we use web as our resource, but there are too many noisy data, how to eliminate them before IR should be studied; 2. NE recognition is important to our system, how to recognize more useful named entities should be researched in the future.

Reference

[1] Yutaka Sasaki, Chuan-Jie Lin, Kuang-hua Chen, Hsin-Hsi Chen. Overview of the NTCIR-6 Cross-Lingual Question Answering (CLQA) Task. Proceedings of the NTCIR6 Workshop, 2007

[2] Teruko Mitamura, Eric Nyberg, Hideki Shirna et al. Overview of the NTCIR-7 ACLIA: Advanced Cross-Lingual Information Access. Proceedings of the NTCIR-7, 2008

[3] http://aclia.lti.cs.cmu.edu/wiki/moin.cgi/Home

[4] Yang, H., Chua, T. FADA: Find All Distinct Answers, in 'Proceedings of WWW Alt. '04', New York, NY, USA, pp. 304-305 2004

[5] Dong-feng Cai, Dong-yuan Li, Yu Bai, Bo Zhou. KECIR Information Retrieval System for NTCIR7 IR4QA Task. Proceedings of the NTCIR-7, 2008

[6] T.Landuaer, S.Dumias. A solution to plato's problem: The latent semantic analysis theory of acquisition. Psychological Review, 104(2) : 211 - 240 1997

[7] Peter D.Turney. Similarity of Semantic Relations. Computational Linguistics, 32(3):379-416 2006

[8] Sanda Harabagiu, Steven Maiorano, Alessandro Moschitti, and Cosmin Bejan. Intentions, implicate-res and processing of complex questions. In Sanda Harabagiu and Finley Lacatusu, editors, HLT-NAACL 2004: Workshop on Pragmatics of Quest-ion Answering, pages 31–42, Boston, Massachu-setts, USA, May 2-May 7 2004