Description of the NTOU Complex QA System

Chuan-Jie Lin Yu-Min Kuo

Department of Computer Science and Engineering
National Taiwan Ocean University

{cjlin, corn.guo}@ntou.edu.tw

ABSTRACT

This paper provides the description of our complex QA system, the NTOU XQA System participated in NTCIR-8 CCLQA Task. This QA system can answer several types of factoid questions and 5 types of complex questions defined in NTCIR-8 CCLQA Task. Different strategies are designed for finding answers to different types of questions. Named entity recognition, distance scores of question keywords, answer information patterns, and search results from the Web are techniques integrated in these strategies. The best F-measure score achieved by our system is 19.88% in monolingual task and 13.62% in cross-lingual task. But unfortunately the official evaluation is incorrect.

Category and Subject Descriptions

H3.4 [Information Storage and Retrieval] Systems and Software - *Question-answering (fact retrieval) systems*

General Terms

Performance, Experimentation

Keywords

Question answering, complex questions

1. INTRODUCTION

Cross-lingual QA has been one of the tracks in NTCIR evaluations since 2005. CLQA tasks in NTCIR-5 and NTCIR-6 [[1]] dealt with factoid questions, including PERSON, LOCATION, TIME, ORGANIZATION, NUMBER, etc. We participated in NTCIR-5 CLQA task and built our first cross-lingual QA system [[2]].

CCLQA task in NTCIR-7 [[3]] dealt with complex questions, including BIOGRAPHY, DEFINITION, RELATIONSHIP, and LIST/EVENT questions. Many research results were presented in this workshop [4][5][6][7][8].

This time, WHY questions are newly added in the formal test set which ask for causes or explanations. Girju *et al*. have examined verbs which denoting causal relations [9] and 沈天佐 *et al*. [10] tested the correctness of different causal patterns in English.

We participated in NTCIR-8 CCLQA Task for Chinese-to-Chinese (C-C) and English-to-Chinese (E-C) tasks. This is our first time developing a QA system for complex questions, although we already have some experiences on handling BIOGRAPHY, DEFINITION, and WHY questions.

This paper is structured as follows. Section 2 gives a description of the architecture of our complex QA system, NTOU XQA System. Section 3 gives details of methods to do answer type classification. Section 4 introduces the procedures to find answers from the Internet for factoid and WHY questions. Section 5 expresses how we extract final answers from the ACLIA2 Corpus. Section 6 presents the strategies used in our formal runs and their evaluation results. Section 7 concludes this paper.

2. SYSTEM DESCRIPTION

The architecture of the NTOU XQA System follows a typical QA procedure. This QA system is designed to find answers from the web pages, not from a static document collection.

Figure 1 shows the architecture of the NTOU XQA System. Given an input question, the system first decides its answer type (including factoid and complex question types) and extracts the focus and keywords of the question. A search engine is used to retrieve relevant web pages from the Internet. Possible answers are mined in the nuggets or the full text of retrieved web pages. Different types of questions will be answered by different answer finding modules. More details will be described in the subsequent sections.

To participate in NTCIR-8 CCLQA Task, our system has to be adapted in order to find answers from a static document collection. For BIOGRAPHY, EVENT, and RELATIONSHIP questions, their answer-finding strategies are applied directly to the relevant documents retrieved from the ACLIA2 Corpus. For DEFINITION, WHY, and factoid questions, possible answers found from the Internet are matched in the relevant documents retrieved from the ACLIA2 Corpus.

Cross-lingual Experiments

Because we have not developed our own cross-lingual handling methods yet, to perform cross-lingual QA from English to Chinese, we submitted all the English question sentences to the Google Translate webpage, and then used the Chinese translation results as the question inputs as in monolingual question answering.

3. ANSWER TYPE CLASSIFICATION

There are factoid and complex questions in the NTCIR-8 CCLQA question set. The methods to guess types of factoid questions and types of complex questions are different in our system.

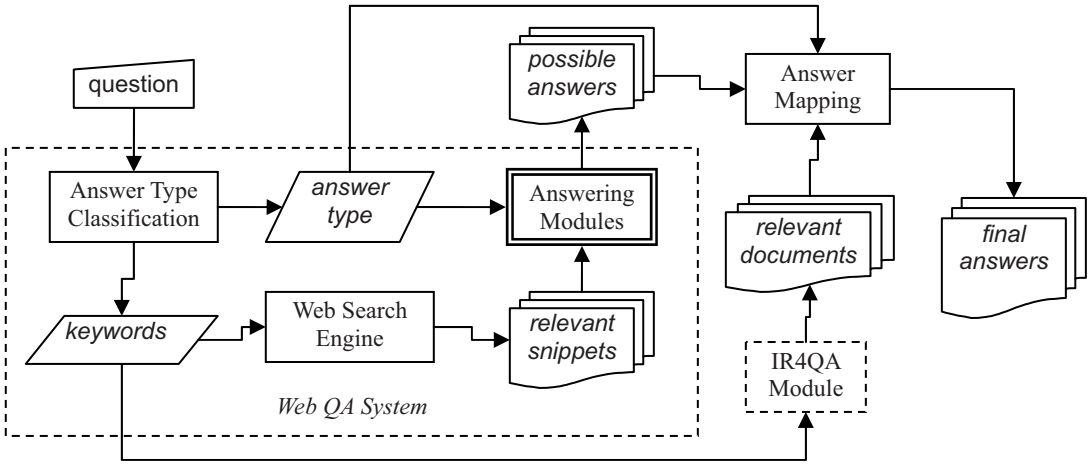

Figure 1. Architecture of the NTOU XQA System

3.1 Rule-based Classification

Lin [11] has developed several hand-crafted rules for detecting factoid questions, including PERSON, LOCATION, OBJECT, TIME, ORGANIZATION, and QUANTITY. We use these rules to detect factoid questions. Samples of these rules are as follows.

person [哪] [幾,**NumSet**] [個,位] {**PersonSet**: 總統,導演...}
location [什麼,甚麼,啥,何] {**LocSet**: 國家,城市,鄉...}
time [什麼,甚麼,啥,何] {**TimeSet**: 年,時候,天...}

Where NumSet is the set of positive integer expressions (whether in Chinese or English, identified by simple rules), PersonSet is the set of hyponyms of "person", LocSet the hyponyms of "location", and TimeSet the hyponyms of "time". (The meanings of the Chinese words seen in the examples above are given here: 哪 "which", 幾 plural when counting, [個,位] legal quantification words for counting persons in Chinese, [什麼,甚麼,啥,何] synonyms of "what".)

Although Lin has also built rules for three kinds of complex questions (WHY, BIOGRAPHY, and DEFINITION), those rules are too simplified that not all the cases can be covered. We augmented old rules and invented news rules for complex questions by observing questions from the NTCIR-7 CCLQA question set. Since WHY-questions were not covered in NTCIR-7 CCLQA task, rules for WHY-questions are the same as Lin's rules. Note that if no rule matches a given question, its answer type will be guessed as OTHER.

3.2 Clue-mining Classification

We also tried automatic methods to learn question classification. *Clue terms* were extracted from the NTCIR-7 CCLQA questions. We tried three kinds of units of clue terms: character bigrams, words, or bi-words. Terms whose frequencies exceeded a pre-defined threshold were selected as clue terms. For each of the four answer types defined in NTCIR-7 CCLQA Task, a set of clue terms were collected. Table 1 lists some clue terms of the four complex answer types.

The meaning of the Chinese words appearing in Table 1 are: 是 "is", 誰 "who", 請 "please", 告訴 "tell", 我 "me", 什麼 "what", 何謂 "what_is", 叫 "so-called", 列出 "list", 的 an auxiliary, 事件 "event", 詳情 "details", 關係 "relationship", 為何 "is_what".

Table 1. Examples of clue terms (in bigram, word, and biword)

AType	BIO	DEF	EVT	REL
Bigram	是誰 誰是 告訴	什麼 麼是 何謂	列出 事件 發生	關係 什麼 之間
AType	BIO	DEF	EVT	REL
Word	誰 是 請	什麼 是 何謂	的 列出 請	關係 的 什麼
AType	BIO	DEF	EVT	REL
Biword	是 誰 誰 是 告訴 我	什麼 是 是 什麼 什麼 叫	的 事件 的 詳情 請 列出	的 關係 什麼 關係 關係 為何

(BIO: BIOGRAPHY; DEF: DEFINITION; EVT: EVENT; REL: RELATIONSHIP)

When a question is submitted, the clue terms are matched in the question. The question's answer type is chosen as the one whose clue terms appear in the question for more times than the clue terms of other types. Take ACLIA2-CT-0012 "What is the relationship between sleep and longevity" as an example to explain how to use clue terms (in words) to guess its answer type:

Word-segmented question:
睡眠 與 壽命 長短 的 關係 為何

Type	Matched clues	Score	Guess
BIO	的	1	
DEF	的 為何	2	
EVT	的	1	
REL	與 的 關係 為何	4	V

4 RELATIONSHIP clue terms appear in the question while there is only one or two clue terms in other types appear in the question. The answer type of this question is decided as RELATIONSHIP.

We also calculate the ***clue-coverage ratio*** of a question which is defined as the number of matched clue terms belonging to the question's answer type divided by the number of terms in this question. If its clue-coverage ratio is smaller than a pre-defined threshold, the answer type of this question will be changed into OTHER.

3.3 Classification Voting

To merge the guesses from these different strategies, a voting mechanism is designed to make the final decision. A factoid type has the highest preference. A complex type is decided by votes from different strategies where BIO > DEF > EVT > REL. The type "OTHER" does not count as a vote. Note that for a WHY-question, it gets no votes from the clue-mining classification methods and can only be detected by Lin's rules. If a question gets no vote at all, it will be classified according to the highest clue-coverage ratio.

3.4 Keyword Extraction

The answer type classification rules also consist of patterns to extract question keywords. When the answer type of a question is determined, its keywords can be prepared at the same time.

For a complex question, its keywords are extracted as follows: remove all the clue terms related to its answer type and collect the remaining substrings as keywords.

Again, we use ACLIA2-CT-0012 as an example to demonstrate how to use clue terms (in words) to extract answer keywords. The matched clue terms in the question sentence are removed (blackened in the example) and the remaining segments (underlined) are keywords.

Word segmented:
睡眠 與 壽命 長短 的 關係 為何
After clue term removed:
睡眠 與 壽命 長短 的 關係 為何

4. ANSWERING FROM THE WEB

As we described in Section 2, for DEFINITION, WHY, and factoid questions, possible answers are first mined from the Internet, and then matched in the relevant documents. The answers to the BIOGRAPHY, RELATIONSHIP, and EVENT questions are searched directly from the relevant documents retrieved from the ACLIA2 Corpus.

The online answer mining methods are described in the following subsections.

4.1 Factoid Questions

The methods of finding answers to factoid questions are exactly the ones used in Lin's work. The following paragraph is a brief description of the procedures of finding answers to a factoid question.

Question keywords are combined as one query and submitted to an online search engine. Answer candidates are identified in the top N nuggets by the help of a named entity recognition system. Each candidate is scored according to the number of nuggets where it occurs and the distance scores related to the question keywords surrounding it in the nuggets. Finally, at most top 10 answers are proposed according to their scores.

The response for a factoid question by our QA system is the exact string of the answer itself together with a set of snippets providing evidences for verification. The evidence snippets are the snippets in the search results which support the same answer.

4.2 DEFINITION and WHY Questions

The methods of finding definition and causal answers from the Internet also follow Lin's work, which are described here.

DEFINITION

There are some common patterns which express definitions in a written text, such as "*XXX* 就是" (*XXX is a ...*) or "*XXX* 的定義是" (*the definition of XXX is ...*). Some patterns are invented to construct queries to the search engine. Snippets matching the given patterns are considered as carrying definitions. For example, given a question "類固醇是什麼" ("What is the so-called steroid"), the following queries are submitted to the search engine:

"類固醇就是" ("*steroid* is")
"類固醇的定義是" ("the definition of *steroid* is")
"類固醇的意思是" ("the meaning of *steroid* is")

Note that the queries are given as phrases so they should not be separated in the retrieved snippets. For a snippet containing the query, the substring directly following the query is extracted as a possible definition. The bordered text in the following example[1] is one possible definition.

類固醇就是治療氣喘抗發炎藥物中最有效的藥物
(Steroid is the most effective anti-inflammatory medicine used to treat asthma)

WHY

Some cause-effect clue words, such as "因為" (*because*) and "因此" (*therefore*), are collected in advance. Each clue word is combined with all the questions keywords and submitted to the search engine. The original web pages in the result lists are downloaded and further analyzed according to its rhetorical structure to find out the causal parts. The causal parts are returned as the exact answer strings and the whole sentences as their evidence snippets.

Take the question "禽流感為什麼會威脅人類" ("Why is the bird flu a threat to humans") as an example. A query "禽流感 威脅 人類 因為" ("the bird flu" "threat" "humans" "because") is constructed for the cause-effect clue word "因為" (*because*) and submitted to the search engine. All three question keywords (marked as bold and italic in the example) appear in the effect passage, so the causal part of the text (the bordered text) is extracted as a possible reason[2].

流感病毒愈像***禽流感***病毒，其對***人類***的***威脅***也愈大，
因為人類的免疫系統過去從未暴露在這些病毒的環境下

(The greater similarity that a flu virus is to the ***bird flu*** virus means a bigger ***threat*** to ***humans***, **because** the humans have never been in an environment with this kind of virus.)

5. ANSWER MATCHING

5.1 Relevant Document Retrieval

In the architecture of our QA system, there should have been an IR module retrieving relevant documents from the ACLIA2 Corpus. We do have our own IR system. However, we failed to complete the indexing of the corpus in time, so we decided to adopt the

[1] Selected from the web page:
http://www.tainantb.gov.tw/?aid=302&page_name=detail&iid=75
[2] Selected from the web page:
http://www.dajiyuan.com/b5/4/2/7/n462106.htm

results from the IR4QA groups. In the future, we will complete the IR module to see the real performance of our QA system.

Relevant documents were selected in a manner of merging the results from the IR4QA groups. At the stage of releasing IR4QA results to CCLQA participants, we downloaded 5 monolingual runs from one IR4QA group and 18 cross-lingual runs from 3 groups. All the retrieved documents in these runs were voted by the groups and runs. A document was ranked first if it was retrieved by most groups. Among those documents that got the same votes from the 4 groups (MLQA runs were treated as contributed from a different group) were ranked according to the votes by the 23 runs. To break a tie, the documents were ranked according to their ranks in the monolingual runs.

5.2 Web Answer Mapping

For DEFINITION, WHY, and factoid questions, the top N answers and their evidence snippets provided by the NTOU XQA System are used to find answers in the relevant documents from the ACLIA2 Corpus. If no answers are found from the Internet, the sentences most similar to the questions are extracted instead.

Factoid

Three scoring functions are defined for mapping Web answers:

$$S_{ans}(s) = \max_{a \in A_{web}} \frac{|s \cap a|}{|a|} \quad (1)$$

$$S_{evd}(s) = \max_{a \in A_{web}, e \in Evd(a)} \frac{|s \cap e|}{|e|} \quad (2)$$

$$S_{qkw}(s) = \frac{|s \cap Q|}{|Q|} \quad (3)$$

where s is the sentence in one of the relevant documents which is going to be scored, A_{web} is the set of Web answers mined from the Internet, $Evd(a)$ is the set of evidence snippets of an Web answer a, and Q is the question sentence.

The lengths of a string and the overlap scores of two strings in the three scoring functions are measured in words. So all the sentences, answers and snippets are word segmented beforehand.

Each sentence in the relevant documents will have three scores according to these functions where S_{ans} measures its similarity to a Web answer, S_{evd} measures its similarity to an evidence snippet of a Web answer, and S_{qkw} measures its similarity to the question sentence.

We do not combine the three scores in linear combination. Instead, we use them as different sorting keys to decide the ranks. Sentences having larger S_{ans} scores are ranked higher. When two sentences have the same S_{ans} scores, the one with a larger S_{qkw} score is ranked higher. And S_{evd} scores are used to be the third sorting key. I.e. the precedence of the ranking procedure is $S_{ans} > S_{qkw} > S_{evd}$.

WHY

The same methods for answering factoid questions are applied to answer WHY questions as well, only that the ranking precedence is different: $S_{ans} > S_{evd} > S_{qkw}$.

An alternative method was experimented in the CCLQA task. We define another scoring function as follows:

$$S_{ckw}(s) = \frac{|s \cap Ckw|}{|s|} \quad (4)$$

where Ckw is the set of cause-effect keywords (e.g. “because” and “therefore”). When using this scoring function, the ranking preference is set to be $S_{ckw} > S_{qkw}$.

DEFINITION

The result snippets from the search engine are compared with the sentences of the relevant documents. Similarity is measured in the Jaccard coefficient of character bigrams. The top N similar sentences are proposed as the final answers.

$$S_{Jcd}(s) = \max_{a \in A_{web}} \frac{|s \cap a|}{|s \cup a|} \quad (5)$$

5.3 Complex Answer Matching

The answer matching methods for different types of complex questions are described as follows.

BIOGRAPHY

We have collected a lot of patterns possibly giving hints of biographical information. Table 2 lists some of the examples of biographical patterns, where <person> denotes the location of a person name and <bioinfo> denotes the biographical information. As we can see, some patterns are used to express more than one type of information, such as the phrase “出生於” (was born in/at) can deliver either the person’s birth date or his birth place.

Table 2. Examples of Biographical Patterns

Biographical Patterns	Possible Types of Biographical Information
<person>（<bioinfo>）	titles, nickname
<person>出生於<bioinfo>	birth date, birth place
<person>出生日期：<bioinfo>	birth date
<person>是<bioinfo>人	birth place

(meaning:: 出生於 “was born in/at”; 出生日期 “birth date”; 是“is”; 人“people of/from”)

Given a biography question, the <person> parts in the patterns are replaced by the person name extracted from the question, and then the patterns are searched in the relevant documents. Sentences which contain any biographical patterns are possible answers. They are ranked by their lengths (in characters). If two sentences have the same lengths, the sentence which comes from a higher-rank document is preferred. However, each document will contribute no more than two sentences as answers in order to exclude similar information.

These biographical patterns were automatically learned from the Web. We extracted thousands of patterns but found that most of them were useless. Moreover, we found that frequent patterns could seldom been matched in the documents in the ACLIA2 Corpus. Currently we have no idea which patterns can be really useful. We will perform a full investigation of the qualities of these patterns in the future.

EVENT

Temporal expressions and location names in relevant documents are identified by a NER system. For each document, we select one sentence which contains at least one temporal expression and

at least one location name. If a newly selected sentence contains the same temporal expression or location name as the ones in the previous selected sentences, it will be discarded.

RELATIONSHIP

As described in Section 3.4, two keyword substrings of a relationship question will be extracted. They are regarded as the two targets in the relationship question. Sentences containing the two targets are possible answers. The distance between the two targets in a sentence is the main ranking key, i.e. the nearer that the two targets are close to each other, the higher this sentence will be ranked.

6. EXPERIMENTAL RESULTS

6.1 Answer Type Classification

Table 3 and Table 4 list the answer type classification performance for monolingual QA. Table 3 lists the numbers of questions being classified in the gold standard and by our system in every answer types, where rows represent the numbers in the gold standard and columns represent the numbers classified by our system. Table 4 lists the recall, precision, and F1-score for each answer type, respectively. The meanings of the symbols used in the tables are as follows:

B	Biography	L	Location
D	Definition	T	Date
E	Event	O	Organization
R	Relationship	Q	Quantity
W	Why	J	Object
P	Person	Atype	Answer type

As we can see in Table 4, the overall accuracy (i.e. the recall in total) of the classification is 70%. For DATE and WHY types, both recall and precision are higher than 90%. Our system failed to detect any organization question, which means that the classification rules need to be re-written.

Table 3. Answer type classification (MLQA)

Gld\Sys	B	D	E	R	W	P	L	T	Q	J	Total
B	**7**			2		1					10
D		**9**								1	10
E	2	3	**4**	7			1		1	2	20
R	1			**19**							20
W		1		1	**18**						20
P						**5**					5
L				1			**3**			1	5
T								**5**			5
O				4		1					5
Total	10	13	4	34	18	7	4	5	1	4	100

Table 4. Answer type evaluation (MLQA)

Atype	Recall	Precision	F1
B	70.00	70.00	70.00
D	90.00	69.23	78.26
E	20.00	100.00	33.33
R	95.00	55.88	70.37
W	90.00	100.00	94.74
P	100.00	71.43	83.33
L	60.00	75.00	66.67
T	100.00	100.00	100.00
O	0.00	0.00	0.00
Total	70.00	73.68	71.79

However, we have different opinions on two questions in the ACLIA Test Set. They are ACLIA2-CT-0056 “Please list the movies in which Zhao Wei participated” and ACLIA2-CT-0057 “Please list the New Year films made by Feng Xiaogang”. They are classified as EVENT questions in the ACLIA Test Set. But to us, their answers are movie titles thus more like ARTIFACT (factoid, classified as OBJECT in our system) questions.

Table 5 and Table 6 list the answer type classification performance for CLQA. Not surprisingly, the overall accuracy (recall in total) drops to 50%.

Table 5. Answer type classification (CLQA)

Gld\Sys	B	D	E	R	W	P	L	T	Q	J	Total
B	**4**	1				5					10
D		**10**									10
E	4	3	**5**	3			1			4	20
R		8	1	**6**			1			4	20
W		3		2	**14**				1		20
P						**5**					5
L				2			**3**				5
T		1		1				**3**			5
O				3		1				1	5
Total	8	26	6	17	14	11	5	3	1	9	100

Table 6. Answer type evaluation (CLQA)

Atype	Recall	Precision	F1
B	40.00	50.00	44.44
D	100.00	38.46	55.56
E	25.00	83.33	38.46
R	30.00	35.29	32.43
W	70.00	100.00	82.35
P	100.00	45.45	62.50
L	60.00	60.00	60.00
T	60.00	100.00	75.00
O	0.00	0.00	0.00
Total	50.00	55.56	52.63

6.2 Run Description

We only used the QUESTION fields in the test collection to produce formal runs. English questions were first translated by Google Translate System and then processed in the same way as in monolingual QA.

We submitted three monolingual runs and three cross-lingual runs according to the same three strategies described as follows.

Run 1

For each factoid question, only the best sentence containing the top-1 answer was proposed. If more than sentences contained the top-1 answer, they were ranked by the following preference: $S_{ans} > S_{qkw} > S_{evd}$. If no exact answer was provided by the NTOU XQA System, we followed the strategy of Run 3 to prepare answers.

For BIOGRAPHY, DEFINITION, EVENT and RELATIONSHIP questions, top 30 sentences were proposed by the methods described in Section 5.3. When comparing temporal expressions in finding EVENT answers, we did not perform temporal resolution, i.e. comparing was performed on the surface strings.

If only a few sentences were found in the first place, a back-off model was performed which selected sentences with a length between 10 to 30 characters containing question keywords in the order of the rankings of the relevant documents. There is no

need to perform the back-off model for DEFINITION questions.

For WHY questions, top 10 sentences scored by the three similarity scores were proposed in the following preference order: $S_{ans} > S_{evd} > S_{qkw}$.

Run2

For each factoid question, the best sentences of the top 10 answers were proposed. We selected one sentence for each web answer. For the sentences containing the same answer, they were ranked by the following preference: $S_{ans} > S_{qkw} > S_{evd}$. When no exact answer was provided by the NTOU XQA System, we followed the strategy of Run 3 to prepare answers.

For WHY questions, the top 10 sentences scored by the occurrences of cause-effect keywords and the question similarity scores were proposed in the following preference order: $S_{ckw} > S_{qkw}$.

For other kinds of complex questions, same strategies were used as those in Run 1.

Run3

For factoid questions, top 10 sentences most similar to the question sentences are chosen as the answers.

For BIOGRAPHY questions, only sentences with lengths shorter than 20 characters were considered. At most top 30 sentences were proposed.

For DEFINITION questions, besides the sentences selected in Run1, the back-off model was also used until totally 30 sentences were proposed.

For EVENT questions, temporal resolution was performed to transform a temporal expression into a normal form before comparing two expressions.

For WHY questions, same strategies as those in Run 2 were used.

6.3 Performance

Table 7 presents the official evaluations of our six runs, where the first three rows are for monolingual runs and the last three rows are for cross-lingual runs. It is obvious that our system proposed too many and too long answers so that the precision was extremely low. It can be improved by only proposing exact answers for factoid questions and propose fewer sentences for complex questions.

Table 7. Official human evaluation

Runs	Recall	Precision	F1
CT_CT_01_T	30.46	8.61	18.15
CT_CT_02_T	**40.93**	**8.63**	**19.88**
CT_CT_03_T	26.72	7.91	14.61
EN_CT_01_T	17.69	5.69	10.96
EN_CT_02_T	**28.99**	**6.18**	**13.62**
EN_CT_03_T	22.62	5.39	11.59

In order to see the performance of the three runs without the answer type classification errors, the data in Table 8, Table 9, and Table 10 are prepared from the correctly classified questions. These three tables present the recall, precision, and F3 scores for every answer types, where X means misclassified questions.

But unfortunately there are errors in the human evaluation data. Because we proposed same answers for some types of complex questions in some runs, the performances in these types should be exactly the same. However, as we can see in the tables, the data are not identical.

We were informed that there was a problem in the XML parsing function used in the evaluation system. Some sentences in the runs were truncated before being saved in the database. This parser was not developed by the organizers and its mistakes were unpredictable.

The organizers released a set of automatic evaluation results for the original (untruncated) runs. Table 11 to Table 13 give the corresponding data prepared in the same way as for Table 8 to Table 10.

We tend not to make conclusions based on these data, because it is still possible that they are not the exact results. We compared the human evaluation files with our submitted runs and found 127 sentences (in 62 topics) which were not assessed in their full lengths. We could not be sure that there was no answer appearing in the truncated strings.

7. CONCLUSION

This year we participated in the NTCIR-8 CCLQA C-C and E-C subtasks. The answer type classification is performed by using rule matching and clue term detection. Questions keywords are extracted at the same time.

For DEFINITION, WHY, and factoid questions, we first used our QA system, NTOU XQA System, to find possible answers from the Internet and then searched them in the ACLIA Corpus to extract sentences containing answers. For BIOPGRAPHT questions, thousands of biographical patterns automatically learned from the Internet were used to search in the ACLIA Corpus. For EVENT and RELATIONSHIP questions, simple strategies were proposed to rank sentences containing questions keywords.

The accuracy of answer type classification is 70% in monolingual runs and 50% in cross-lingual runs. There is still space to improve the performance.

The best F3 score achieved by our system is 19.88% in monolingual task and 13.62% in cross-lingual task. But unfortunately there was a truncation problem during the evaluation procedure, the evaluation results are not correct.

8. REFERENCE

[1] Sasaki, Y., Lin, CJ., Chen, KH., and Chen, HH. 2007. Overview of the NTCIR-6 Cross-Lingual Question Answering (CLQA) Task. In *Proceedings of NTCIR-6*, 153-163.

[2] Lin, CJ., Tzeng, YC., and Chen HH. 2005. System Description of NTOUA Group in CLQA1. In *Proceedings of NTCIR-5*, 250-255.

[3] Mitamura, T., Nyberg, E., Shima, H., Kato, T., Mori, T., Lin, CY., Song, R., Lin, CJ., Sakai, T., Ji, D., and Kando N. 2008. Overview of the NTCIR-7 ACLIA Tasks: Advanced Cross-Lingual Information Access. In *Proceedings of NTCIR-7*, 11-25.

[4] Wu, Y., Chen W., and Kashioka H. 2008. NiCT/ATR in NTCIR-7 CCLQA Track: Answering Complex Cross-lingual Questions. In *Proceedings of NTCIR-7*, 26-32.

[5] Shima, H., Lao, N., Nyberg E., and Mitamura T. 2008. Complex Cross-lingual Question Answering as a Sequential Classification and Multi-Document Summarization Task. In *Proceedings of NTCIR-7*, 33-40.

[6] Mori, T., Okubo T., and Ishioroshi M. 2008. 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. In *Proceedings of NTCIR-7*, 41-48.

[7] Ren, H., Ji, D., He, Y., Teng C., and Wan J. 2008. Multi-Strategy Question Answering System for NTCIR-7 C-C Task. In *Proceedings of NTCIR-7*, 49-53.

[8] Lee, YH., Lee, CW., Sung, CL., Tzou, MT., Wang, CC., Liu, SH., Shih, CW., Yang RY., and Hsu, WL. 2008. Complex Question Answering with ASQA at NTCIR 7 ACLIA. In *Proceedings of NTCIR-7*, 70-76.

[9] Girju, R. and Moldovan, D. 2002. Text Mining for Causal Relations. In *Proceeding of FLAIRS Conference 2002*, 360-364.

[10] 沈天佐，林川傑，陳信希. 2003. 以網際網路內容為基礎之問答系統 “Why” 問題之研究. In *Proceedings of ROCLING-15*, 211-229, written in Chinese.

[11] Lin, CJ. 2004. *A Study on Chinese Open-Domain Question Answering System*. Ph.D. dissertation, National Taiwan University.

Table 8. Recall of correctly classified questions

Runs	B	D	E	R	W	P	L	T	X	Total
CT_CT_01_T	35.74	18.99	**24.41**	46.26	15.13	36.36	33.33	20.00	33.15	30.46
CT_CT_02_T	**37.80**	18.99	23.34	46.26	**25.77**	**80.00**	**100.00**	**84.85**	**36.58**	**40.93**
CT_CT_03_T	34.83	7.67	12.20	**46.43**	25.07	20.00	**100.00**	0.00	27.30	26.72
EN_CT_01_T	23.59	10.03	20.74	34.64	7.32	16.36	33.33	0.00	19.56	17.69
EN_CT_02_T	23.59	**19.45**	23.10	42.83	21.31	**80.00**	66.67	82.35	21.84	28.99
EN_CT_03_T	22.40	10.03	22.41	34.64	21.77	20.00	**100.00**	0.00	20.95	22.62

Table 9. Precision of correctly classified questions

Runs	B	D	E	R	W	P	L	T	X	Total
CT_CT_01_T	4.56	12.88	10.55	7.48	5.01	**3.91**	**4.77**	**6.06**	**12.48**	8.61
CT_CT_02_T	**4.90**	**12.90**	10.31	7.54	**8.99**	1.50	0.93	5.13	11.02	**8.63**
CT_CT_03_T	4.36	4.97	**11.66**	7.29	7.37	0.28	1.23	0.00	10.27	7.91
EN_CT_01_T	3.25	5.99	7.30	14.21	2.23	3.36	**4.77**	0.00	6.23	5.69
EN_CT_02_T	2.48	12.68	8.27	**17.53**	6.75	1.58	0.74	4.09	4.35	6.18
EN_CT_03_T	3.14	6.08	7.53	14.70	7.01	0.28	1.10	0.00	4.73	5.39

Table 10. F3 of correctly classified questions

Runs	B	D	E	R	W	P	L	T	X	Total
CT_CT_01_T	21.11	17.62	**21.39**	24.86	12.18	**18.57**	**20.85**	16.26	16.52	18.15
CT_CT_02_T	**22.48**	17.62	20.57	25.02	**19.51**	12.50	8.49	**29.29**	**17.62**	**19.88**
CT_CT_03_T	20.50	7.19	11.28	**28.78**	17.93	2.48	11.08	0.00	15.63	14.61
EN_CT_01_T	14.51	9.10	15.72	26.93	5.79	11.80	**20.85**	0.00	10.08	10.96
EN_CT_02_T	12.67	**17.80**	17.44	28.62	14.74	12.86	6.73	26.66	10.07	13.62
EN_CT_03_T	13.88	9.10	16.84	27.28	15.32	2.49	9.99	0.00	10.16	11.59

Table 11. Recall of correctly classified questions (untruncated runs, auto evaluation)

Runs	B	D	E	R	W	P	L	T	X	Total
CT_CT_01_T	65.83	33.33	26.64	30.60	66.67	56.36	69.05	42.25	46.90	46.15
CT_CT_02_T	65.83	100.00	26.64	47.75	100.00	80.00	69.05	33.30	49.68	51.33
CT_CT_03_T	64.04	15.69	11.65	20.92	100.00	43.64	51.90	27.49	34.89	34.83
EN_CT_01_T	28.29	20.00	17.80	21.25	66.67	16.36	28.38	30.66	26.96	27.25
EN_CT_02_T	41.13	60.00	27.42	33.92	66.67	80.00	33.42	30.63	30.85	36.82
EN_CT_03_T	36.76	9.41	17.80	21.25	100.00	43.64	28.38	30.01	29.37	30.28

Table 12. Precision of correctly classified questions (untruncated runs, auto evaluation)

Runs	B	D	E	R	W	P	L	T	X	Total
CT_CT_01_T	6.45	10.10	18.61	16.45	7.54	7.14	7.24	9.91	14.27	12.78
CT_CT_02_T	6.45	5.51	18.61	20.52	0.69	0.97	7.24	9.47	12.28	11.28
CT_CT_03_T	7.27	1.37	7.23	17.86	0.98	0.85	19.43	9.21	10.10	9.52
EN_CT_01_T	10.07	3.40	7.73	15.34	6.34	3.36	9.00	9.30	7.05	8.04
EN_CT_02_T	9.14	3.65	14.12	17.14	0.50	1.07	7.77	9.19	9.30	8.77
EN_CT_03_T	6.62	0.82	7.73	15.34	0.89	0.85	9.00	9.06	7.16	7.37

Table 13. F3 of correctly classified questions (untruncated runs, auto evaluation)

Runs	B	D	E	R	W	P	L	T	X	Total
CT_CT_01_T	34.01	27.10	24.75	26.13	36.68	33.00	35.77	28.78	26.71	28.23
CT_CT_02_T	34.01	33.33	24.75	39.11	6.50	8.59	35.77	23.20	25.97	25.79
CT_CT_03_T	35.81	7.67	10.58	19.18	9.03	6.91	33.15	21.31	19.17	18.85
EN_CT_01_T	16.96	13.44	14.49	20.30	31.87	11.80	22.07	21.72	15.01	18.13
EN_CT_02_T	26.11	21.34	23.35	30.44	4.70	9.36	23.16	21.99	17.24	20.35
EN_CT_03_T	23.54	4.60	14.49	20.30	8.24	6.91	22.07	21.50	15.07	17.17