Overview of the NTCIR-6 Cross-Lingual Question Answering (CLQA) Task

Yutaka Sasaki
University of Manchester
yutaka.sasaki@manchester.ac.uk

Chuan-Jie Lin
National Taiwan Ocean University
cjlin@ntou.edu.tw

Kuang-hua Chen Hsin-Hsi Chen
National Taiwan University
khchen@ntu.edu.tw, hh_chen@csie.ntu.edu.tw

Abstract

This paper describes an overview of the NTCIR-6 Cross-Lingual Question Answering (CLQA) Task, an evaluation campaign for Cross-Lingual Question Answering technology. In NTCIR-5, the first CLQA task targeting Chinese, English, and Japanese languages was carried out. Following the success of NTCIR-5 CLQA, NTCIR-6 hosted the second campaign on the CLQA task. Since the handling of Named Entities is a major issue in CLQA, we aimed to promote research on cross-lingual Question Answering technology capable of Named Entities in East Asian languages. We conducted evaluations of seven subtasks: E-J, J-J, J-E, E-C, C-C, C-E, and E-E subtask, where C, E, and J stand for Chinese, English, and Japanese, respectively, and X-Y indicates that questions are given in language X and answers are extracted from documents written in language Y. For the purpose of system development, we provided participants with the sample question/answer pairs and the formal run question/answer pairs used at the previous CLQA task. The Formal Run evaluation was conducted during November 1-7, 2006 with 200 and 150 test questions for Japanese-related CLQA and Chinese-related CLQA, respectively. As a result, 12 research groups world-wide participated in CLQA, and 91 runs were submitted in total.

1. Introduction

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

An evaluation of English Question Answering has been measured in the Question Answering Track at TREC [5]. Japanese QA has been evaluated in NTCIR QAC. Multilingual QA for European Languages has been studies in CLEF Multilingual Question Answering Track [1].

From the linguistic perspective, Cross-Lingual QA is a much more complicated challenge. As a result, organizers decided to keep the direction of the first CLQA task [3] which targeted QA for Named Entities.

On the other hand, we expect high quality QA. That is, top 5 scores are around 90% in the future.

2. Overview

As the second CLQA attempt in NTCIR, we conducted an evaluation of seven subtasks: E-J, J-J, J-E, E-C, C-C, C-E and E-E subtask, where C, E, and J stand for Chinese, English, and Japanese, respectively, and X-Y indicates that questions are given in language X (source language) and answers are extracted from documents written in language Y (target language). Note that an evaluation corresponding to the generic QA for Japanese language was separately conducted in NTCIR QAC.

Target documents

(Corpus for Training)
1. Chinese Dataset (traditional)

CIRB040: United Daily News, United Express, Min Sheng Daily, Economic Daily News 2000-2001

2. Japanese Dataset

Mainichi Newspaper Article Data 2000 – 2001
Yomiuri Newspaper Article Data 2000 – 2001

3. English Dataset

Daily Yomiuri 2000-2001

(Test)
1. Chinese Dataset (traditional)

CIRB020: United Daily News, Economic Daily News, Min Sheng Daily, United Evening News, Star News 1998-1999

2. Japanese Dataset

Mainichi Newspaper Article Data 1998 – 1999

3. English Dataset
EIRB010: Taiwan News; China Times English News 1998-1999
Mainichi Daily News 1998-1999
Korea Times 1998-1999
Hong Kong Standard 1998-1999Corpora

Scope of answers

Each question has only one answer or no answer. Answers are restricted to Named Entities: proper nouns, such as the name of a person, an organization, various artifacts, and numerical expressions, such as money, size, date, etc.

Defining NEs is a very heavy task. So, we use the conventional one for Japanese. Japanese NEs were clearly defined in the NE task of the IREX Workshop [4] (http://nlp.cs.nyu.edu/irex/index-e.html)

The NEs defined by IREX are:

- PERSON
- LOCATION
- ORGANIZATION
- ARTIFACT (product name, book title, law, ...)
- DATE
- TIME
- MONEY
- PERCENT

We adopt these NEs plus NUMEX for CLQA:

- NUMEX

We introduced NUMEX to cover various kinds of numerical expressions other than MONEY and PERCENT.

(Exception 1)

We allow an expression of approximation to be included in answers, such as "about 10" and "more than three" following NTCIR QAC. Basically, the definition of Chinese and English NEs followed the suite.

Question construction

Question construction is the most tricky and difficult part in carrying out an evaluation campaign on CLQA. It is ideal that questions and their answers are all parallel between three languages, Chinese, English, and Japanese. However, it is not easy to find news articles that report the same topics in the three languages. Even if we can find such articles, most of the topics are big news and could be easily predicted by task participants.

We improved the question creation method from NTCIR-5 CLQA to NTCIR-6 CLQA as follows:

(A) Japanese-related CLQA

(NTCIR-5 CLQA)

- Since the Daily Yomiuri articles are English translations of Yomiuri Shimbun articles, we first manually selected corresponding articles between the two corpora, then created an English question by reading an article of the Daily Yomiuri. A Japanese question of the English question was created by referring to the corresponding Japanese article. Thanks to this process, the question/answer pairs of JE and EJ subtasks are parallel.

(NTCIR-6 CLQA)

- A drawback of the above approach was that the questions were strongly affected by written styles of news articles. Therefore, in NTCIR-6, we decided not to enjoy the parallel articles between Japanese and English. First we created 300 English questions and answers by referring to the English corpus, regardless whether or not the articles are English translation of Japanese Newspaper. Then, we searched Yomiuri Shimbun articles containing the same topic as the 300 English questions and created 200 Japanese questions and answers. Since Japanese test corpora is Mainichi Shimbun, we again search articles that can answer the Japanese questions in Mainichi. The Japanese questions that found in Mainichi were used in the formal run. We created new questions and answers by refereeing to the English corpus and Mainichi Shimbun. Thanks to this process, Japanese questions do not resemble the sentences in Japanese test corpus.

(B) Chinese-related CLQA

(NTCIR-5 CLQA)

- C-E subtask: Chinese questions of the C-E subtask are Chinese translations of the English questions for the E-J subtask. This is because CLQA employed the Daily Yomiuri as an English knowledge source.

- C-C and E-C subtasks: Chinese question/answer pairs were created in two different ways: one set was created from the topics of CLIR in NTCIR-5; the other set was created from a real log of an online Chinese QA system (http://nlg.csie.ntu.edu.tw/) [2] with filtering out non-NE questions, and questions which seemed not to have answers in the UDN collection. (It was decided by roughly searching UDN articles by question creators.) And then, Chinese questions were translated into English for EC subtask.

(NTCIR-6 CLQA)

- Participants of CE subtasks in NTCIR-5 CLQA Track suffered from the named entity translation problem: names in the Chinese questions are often from Japanese, not from Chinese or English.

Besides, in order to create comparable subtasks, we decided to create comparable question sets for the four Chinese-related subtasks. First, more about 150 questions and answers were created by referring to the English and Chinese corpora, respectively (hence collecting more than 300 questions). Then, English questions were translated into Chinese, and vise versa. After that, we took the translated questions as queries and employed an IR system to retrieve top relevant documents in the corresponding corpus in order to check the occurrences of correct answers.

Because the corpora adopted in the Chinese-related subtasks are not parallel, less than two-thirds of the questions have answers in both corpora. Furthermore, the types of these questions are not uniformly distributed. As shown in Table 1, most of the questions belong to PERSON or DATE types. After removing some questions of PERSON or DATE types and borrowing some questions from Japanese-related subtasks, 150 questions were collected as the question set of the formal run. Unfortunately, we still cannot avoid the problem of question type distribution. In order to create a more balanced question set next year, we should create double or triple of the numbers of questions for some types.

Training and development sets

We provided participants with the 300 sample question/answer pairs for the E-J, J-E, and C-E subtasks and 200 pairs for the C-C and E-C subtasks. We also provided participants with the 200 question/answer pairs of the previous formal run.

Test sets

For the Formal Run evaluation, we provided 200 questions for each Japanese-related subtask and 150 questions for each Chinese-related subtask. Table 1 shows the number of questions for each question type.

Table 1. Question type distribution of formal run questions

	E-J/J-J/J-E	E-C/C-C/C-E/E-E
ARTIFACT	20	7
DATE	31	39
LOCATION	31	16
MONEY	13	8
NUMEX	20	11
ORGANIZATION	20	16
PERCENT	15	4
PERSON	35	47
TIME	15	2
Total	200	150

Schedule

The time schedule of CLQA was as follows:

- Feb 28, 2006: NTCIR-6 CLQA website established
- April 2006: Call for Participation/ Registration
- mid May 2006: Registration due
- June 2006: Document delivery
- Nov. 1-7, 2006: Formal Run
- Dec. 3-5, 2006: Automatic evaluation results (J-related) and manual evaluation results (C-related) released
- Dec 18, 2006: Manual evaluation results release (J-related)
- March 1, 2007: System Paper submission due
- May 15-18, 2007:NTCIR-6 Workshop Meeting

3. Participants

In total, 12 groups participated in the NTCIR-6 CLQA task, 8 for Chinese-related subtasks and 5 for Japanese-related subtasks. Table 2 and Table 3 show the numbers of runs submitted by participants.

We accepted submissions of at most three official runs and unlimited unofficial runs for each subtask. The numbers of submitted official and unofficial runs are shown separately in the tables.

4. Task definitions

The task definitions are exactly the same as the previous NTCIR CLQA.

Table 2. Number of runs submitted by participants in Japanese-related subtasks

Group	E-J		J-J		J-E	
	official	unofficial	official	unofficial	official	unofficial
Forst	3	3	1	1		
HARAD			1	1		
LTI	3	2		2		
TITLF	3		3			
TTH	3	3	3	3	1	1
Total	12	8	8	7	1	1

Table 3. Number of runs submitted by participants in Chinese-related subtasks

Group	E-C		C-C		C-E		E-E	
	official	unofficial	official	unofficial	official	unofficial	official	unofficial
IASL	3	4	3	7				
ICDCU			1					
ILS	1							
LTI	3	2	3	2				
MHC06	2		2				3	1
NCUTW	1		1					
pircs	3	3	3	3				
WMMKS	1		1			1		
Total	14	9	14	12	0	1	3	1

QA specification

Each question has only one answer or no answer. Answers are restricted to named entities: proper nouns, such as the name of a person, an organization, various artifacts, and numerical expressions, such as money, size, date, etc.

Data specification

In CLQA, the character encoding of the input was BIG5 for Chinese, US-ASCII for English, and EUC-JP for Japanese. The input format of CLQA is defined as follows.

[QID]: "[Question]"

[QID] is the form of [QuestionSetID]-[Lang]-[QuestionNo]-[SubQuestionNo].
[QuestionSetID] is "CLQA2".
[Lang] is one of JA, ZH, and EN.
[QuestionNo] and [SubQuestionNo] consist of four numeric characters starting with "S" or "T" and two numeric characters, respectively. ("S" is for sample questions and "T" for test questions.)
[Question] is a character string.

Example:
CLQA-EN-S0001-00: "When Queen Victoria died?"

The character encoding of the output was BIG5 for Chinese, US-ASCII for English and EUC-JP for Japanese. CLQA defined the following output format.

[QID],[Lang](,"[Answer]",[ArticleID],[Reserved],[Reserved])*

[QID] is the same as in the question file format above. It must be unique in the file, and ordered identically within the corresponding question file. It is, however, allowed that some of the [QID]s are not listed in the file.
[Lang] is one of JA, ZH, and EN.
[Answer] is the answer to the question, and a character string.
[ArticleID] is the identifier of the article or one of the articles used in the process of deriving the answer. The value of the <DOCNO> tag is used for the identifier.
[Reserved] is a field for the future use.

(Example)

CLQA-EN-S0001-00, EN, "1901", ENY-20001101CYM0398, ,
CLQA-EN-S0001-00, JA, "１９０１年", JAY-20001101CYM0398, , , "一九〇一年", JAY-20001101CYM0398, ,

Considering language scalability, the test collection, *i.e.*, a set of gold standard files, is encoded in UTF-8. The format of the test collection for CLQA is defined as follows:

```
<?xml version="1.0" encoding="UTF-8"?>
<QASET>
<VERSION>[Version]</VERSION>

<QA>
```

```
<QUESTION>
<QTYPE>[QType]</QTYPE>
<Q LANG="[Lang]" QID="[QID]">[Question]</Q>
...
</QUESTION>
<ANSWER>
<A LANG="[Lang]" DOCNO="[ArticleID]" GID="[GID]">[Answer] </A>
...
</ANSWER>
</QA>
...
</QASET>
```

[Version] is the version information.
[QID] is the same as in the question file format above.
[Lang] is one of JA, ZH, and EN.
[QType] is one of PERSON, ORGANIZATION, LOCATION, ARTIFACT, DATE, TIME, PERCENT, MONEY, and NUMEX for CLQA.
[Question] is a series of characters.
[ArticleID] is the identifier of the article or one of the articles used in the process of deriving the answer. The value of the <DOCNO> tag is used for the identifier.
[GID] is the group ID (0,1,2,...). This is prepared for evaluating the recall/precision of an answer list but the evaluation of answer lists is out of the scope of CLQA2. If the group number is omitted, it is considered as the group 0.
[Answer] is the answer to the question, and a series of characters. "NIL" means no answer.

(Example)

```
<?xml version="1.0" encoding="UTF-8"?>
<QASET>
<VERSION>NTCIR-5 CLQA …</VERSION>
<QA>
<QUESTION>
<Q LANG="EN" QID="CLQA-EN-S0001-00">When Queen Victoria died?</Q>
<Q LANG="JA" QID="CLQA-JA-S0001-00">ビクトリア女王が亡くなったのはいつ？</Q>
<QTYPE>DATE</QTYPE>
</QUESTION>
<ANSWER>
<A LANG="EN" DOCNO="ENY-20001101CYM0398" GID="0">1901 </A>
...
```

Answer translation

The initial setting of the cross-lingual QA task is to find answers in a different language and then translate them back to the source language. However, the ability to find correct answers will be the major concern of this task. The ability to translate answers back to the source language will be a future evaluation in later CLQAs. Participants were requested to submit answer strings in their original languages (*i.e.*, target languages) in official runs.

RunID format

Regarding official runs, each group was able to submit at most three runs in each subtask. In each official run, only one answer response for each question could be proposed. All of the official runs were assessed. The RunID is an identity for each run and its format is as follows.

GRPID-SL-TL-PfrNo

Here, GRPID is the group ID, SL is the source language of the subtask, and TL is the target language of the subtask; PfrNo is a 2-digit number, which denotes the preference for assessment among the results submitted by the same groups. At most three runs were submitted based on the participants' judgment with the "01", "02", and "03" preference. In the SL and TL columns, 'E' denotes English, 'J' denotes Japanese, and 'C' denotes Chinese. For example, say a group, LIPS, submitted three official runs for the CE subtask. They should be assigned RunIDs of LIPS-C-E-01, LIPS-C-E-02, and LIPS-C-E-03. Following the format described in the Answer Format section, because only one answer response can be proposed, there should be at most one [Answer] string in each line, such as:

```
CLQA-EN-T0001-00, EN, "1901", ENY-20001101CYM0398, ,
CLQA-EN-T0002-00, EN
CLQA-EN-T0003-00, EN, "John Doe", ENY-20010425E1TDY03D000030, ,
```

In order to enlarge the pool size and enable the automatic assessment, we encouraged all participants to submit more results as unofficial runs, *i.e.*, the more the better. In each unofficial run, at most five answer responses for each question can be proposed. The following format was used to name an unofficial run:

GRPID-SL-TL-u-PfrNo

Here, "-u-" is added in the name to denote "unofficial", and other fields have the same meanings as in the format of names of official runs. The amount of unofficial runs which can be assessed will depend on the allowance of time and effort. Since that at most five responses could be proposed, each line in an unofficial run should look like:

```
CLQA-EN-T0001-00, EN, "1901", ENY-20001101CYM0398, , , "1900", ENY-20010724E1TDY02D000050, , , "1998", ENY-20001101CYM0398, ,
```

In the case of submitting translated answers, the same format was used by specifying the language of the answer itself. For example, to submit a response in EC subtask, which intends to find answers of

English questions in Chinese documents, the output for the same question looks like:

CLQA-EN-T0003-00, ZH, "張三", mhn_xxx_20010808_1034915, , CLQA-EN-T0003-00, EN, "John Doe", mhn_xxx_20010808_1034915, ,

The first line is an answer in the target language, and the second line is its translation.

Technique description

In addition to search results, each participating group submitted a file with the filename "GRPID-TechDesc", which was a concise technique description for each submitted run. As mentioned above, GRPID is the group ID. In general, this file should contain the following information.

RunID: as explained in the RunID Section.
IndexUnit: character, bi-character, bi-word, phrase, etc.
IndexTech: the techniques used to process index terms,
e.g., morphology, stemming, POS, etc.
IndexStruc: inverted file, signature file, PAT, etc.
QueryUnit: character, word, phrase, etc.
IRModel: vector space model, probabilistic model, etc.
Ranking: ranking factor for measuring each term, e.g., tf, tf/idf, mutual information, word association, document length, etc.
QueryExpan: techniques used to expand query or no query expansion
TransTech: the translation technique used to deal with cross-language information retrieval, e.g., dictionary-based, corpus-based, MT, etc. The more detailed the information the better, e.g., select-all, select-top-N, WSD, etc.

5. Evaluation method

Each answer response [Answer, DOCNO] was judged. There are three scores used in evaluation:

Right (R): the answer is correct, and the document where it is from supports it.

Unsupported (U): the answer is correct, but the document where it is from cannot support it as a correct answer. That is, there is no sufficient information in the document for users to confirm by themselves that the answer is a correct one.

Wrong (W): the answer is incorrect. Note that even if a substring of an answer response is provided as a correct answer, it will not be judged as a correct one. The same is true for an answer response which is a substring of a real answer.

The assessment of the runs of J-E/E-J/J-J subtasks was conducted independent of the organizers by a Japanese company specializing foreign language communication.

The assessment of E-C/C-C/C-E/E-E subtasks was conducted as follows. Each of the [answer, docID] pairs proposed by the participants (in official or unofficial runs) was judged by three assessors. Majority was taken as its score. If three assessors scored them differently, the organizer would do the final judgment.

Evaluation results were scored by using the accuracy for official runs, and MRR and Top5 scores for unofficial runs.

Accuracy (Top1): is the rate of questions which top 1 answers are correct.

MRR (Mean Reciprocal Rank): is the average reciprocal rank (1/*n*) of the highest rank *n* of a correct answer for each question.

Top5: shows the rate at which at least one correct answer is included in the top 5 answers.

6. Evaluation results

6.1. Results of J-E/J-J/E-J subtasks

Tables 4 show the evaluation results of J-E/E-J/J-J subtasks.

Table 4. Japanese-related CLQA accuracy

Run ID	Right	Right + Unsupported
Forst-E-J-01	0.175	0.195
Forst-E-J-02	0.170	0.180
Forst-E-J-03	0.170	0.180
Forst-J-J-01	0.310	0.335
HARAD-J-J-01	0.085	0.110
LTI-E-J-01	0.095	0.115
LTI-E-J-02	0.095	0.115
LTI-E-J-03	0.070	0.115
LTI-J-J-u-01	0.335	0.360
LTI-J-J-u-02	0.250	0.320
TITFL-E-J-01	0.020	0.040
TITFL-E-J-02	0.030	0.065
TITFL-E-J-03	0.030	0.060
TITFL-J-J-01	0.155	0.190
TITFL-J-J-02	0.130	0.160
TTH-E-J-01	0.130	0.165
TTH-E-J-02	0.120	0.155
TTH-E-J-03	0.070	0.100
TTH-J-J-01	0.245	0.270
TTH-J-J-02	0.235	0.260
TTH-J-J-03	0.270	0.295

In E-J subtask, 12 official runs and 8 unofficial runs were submitted from 4 groups. The best official run was submitted by Forst group. The accuracy was 17.5% with counting only supported answers. It rises to 19.5% if unsupported answers were considered correct as well. The best official runs of the previous E-J subtask were 12.5% and 15.5% for with and without counting unsupported answers, respectively. There found about 35% improvements in the accuracy of the top system.

In J-J subtask, 8 official runs and 7 unofficial runs were submitted from 5 groups. The best official run was submitted by LTI group. The accuracy was 33.5% with counting only supported answers. It rises to 36.0% if unsupported answers were considered correct. Due to the limitation of evaluation resources, unofficial runs of J-E/E-J subtasks were not able to be evaluated except for LTI's run that is the only their submission to J-J subtask.

Only one run was submitted to J-E subtask and no submission to the E-E subtask linked to J-E subtask.

6.2. Results of E-C/C-C/C-E/E-E subtasks

Table 5 shows the best scores of each group in the E-C/C-C/C-E/E-E subtasks.

In C-C subtask, 14 official runs and 12 unofficial runs were submitted from 7 groups. The best official run was submitted by IASL group which accuracy is 49.3% if only supported answers are considered. Its accuracy rises to 52.7% if including unsupported answers. The accuracy of the best run of the previous was 37.5% by the same group. It means a 31.5% improvement in the monolingual Chinese QA.

In E-C subtask, 14 official runs and 9 unofficial runs were submitted from 7 groups. The best official run was submitted by PIRCS group which accuracy is 25.3% if only supported answers are considered. Its accuracy rises to 28.0% if including unsupported answers. The accuracy of the best run of the previous was 12.5% by the same group. It means a 102.4% improvement in the E-C cross-lingual QA!

In C-E subtask, only one run was submitted. But unfortunately the submitting group searched the wrong corpus. Therefore this run was treated as unofficial run and not judged.

In E-E subtask, 3 official runs and 1 unofficial run were submitted by 1 group. The accuracy of the best official run is 18.7% if only supported answers are considered. Its accuracy rises to 20.7% if including unsupported answers. It is the first time to hold an E-E subtask.

Table 5. Chinese-related CLQA accuracy

Run ID	Right	Right + Unsupported
C−C		
IASL	0.520	0.553
ICDCU	0.287	0.340
LTI	0.253	0.260
MHC	0.187	0.213
NCUTW	0.087	0.113
pircs	0.420	0.447
WMMKS	0.133	0.153
E−C		
IASL	0.253	0.340
ILS	0.093	0.107
LTI	0.147	0.200
MHC	0.040	0.073
NCUTW	0.000	0.040
pircs	0.253	0.280
WMMKS	0.053	0.067
E−E		
MHC	0.187	0.207

6.3. Analysis results

(Japanese-related CLQA)

A comparison between E-J and J-J subtasks indicates a difficulty in developing CLQA systems. The performance of E-J CLQA systems sharply drops down to nearly the half of the quality of J-J CLQA systems. Since questions and answers are parallel in E-J and J-J formal runs, contents the questions are the same.

(Chinese-related CLQA)

Comparing the performance of the official runs submitted to the C-C and E-C subtasks by the same group, different groups encounter different level of dropping in the performance. LTI and PIRCS maintain about 60% accuracy comparing to the monolingual task, IASL and WMMKS maintain 40% to 47.2% accuracy, and other groups drop to less than 22%.

No comparison can be made between C-E and E-E subtasks because no groups submitted to both subtasks.

7. Discussion

7.1. Japanese-related CLQA

Two similar questions are intentionally given in Japanese related CLQA subtasks: E-J and J-J subtasks.

QID: T0054/T1054
What is Japan's unemployment rate for May of 1998?
日本の１９９８年５月の失業率は何%ですか？
QID: T0123/T1123

What was the Japan's jobless rate in May 1998?
日本における１９９８年５月の失業率は何%でしたか？

Here, two synonymous English expressions "unemployment rat" and "jobless rate" were used for the translation of "失業率" Two Japanese questions also have slightly different expression "日本の" and "日本における". Interestingly, no system could find a correct answer for T1123 where as official runs of Forst and an unofficial run of LTI could spot correct answer "4.1%" for T1054. This could be "jobless rate" was harder to translate than "unemployment rate". More interestingly, no system could answer these questions in the J-J monolingual setting.

7.2. Chinese-related CLQA

The ability to identify named entities is important in QA. Most teams can extract popular named entities like person, location, and organization names. Teams who did not develop NE identifiers for other named entity types failed to answer questions of these types, as shown in Table 6 and Table 7. It is also true for temporal and numerical expressions.

If we only focus on questions of the types PERSON, ORGANIZATION, LOCATION, and DATE (the four largest subsets), we can see that modules other than NE identification also play important roles which make some systems more effective.

Table 6. Accuracy (%) by question types (CC)

QType	#Q	IASL	ICDCU	LTI	MHC	NCUTW	pircs	WMMKS
ARTI	7	28.57	14.29	14.29	0.00	0.00	42.86	0.00
DATE	39	43.59	33.33	25.64	25.64	0.00	46.15	5.13
LOC	16	87.50	50.00	31.25	18.75	25.00	50.00	37.50
MONY	8	12.50	0.00	12.50	12.50	0.00	25.00	0.00
NUMX	11	27.27	0.00	27.27	0.00	0.00	0.00	0.00
ORG	16	56.25	31.25	18.75	12.50	18.75	31.25	12.50
PRCT	4	25.00	0.00	0.00	0.00	0.00	0.00	0.00
PRSN	47	65.96	34.04	31.91	25.53	12.77	57.45	21.28
TIME	2	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Total	150	52.00	28.67	25.33	18.67	8.67	42.00	13.33

Table 7. Accuracy (%) by question types (EC)

QType	#Q	IASL	ILS	LTI	MHC	NCUTW	pircs	WMMKS
ARTI	7	28.57	0.00	0.00	0.00	0.00	0.00	0.00
DATE	39	15.38	10.26	20.51	7.69	0.00	41.03	2.56
LOC	16	37.50	6.25	18.75	0.00	0.00	43.75	12.50
MONY	8	0.00	12.50	0.00	0.00	0.00	37.50	0.00
NUMX	11	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ORG	16	25.00	12.50	18.75	0.00	0.00	31.25	0.00
PRCT	4	0.00	0.00	0.00	0.00	0.00	0.00	0.00
PRSN	47	42.55	12.77	17.02	6.38	0.00	10.64	10.64
TIME	2	0.00	0.00	0.00	0.00	0.00	100.0	0.00
Total	150	25.33	9.33	14.67	4.00	0.00	25.33	5.33

For EC-subtask, most of the participants used available translation systems to translate English questions into Chinese, and then proceeded as doing monolingual QA. Translation of unknown words especially for named entities became a crucial problem in CLQA.

The effect of translation is not always negative. The question sets of EC and CC subtasks are made parallel (thus questions with QID's same in the last three digits are translations of each other) in order to compare works in MLQA and CLQA. It is interesting that some questions can be answered correctly in CLQA but not in MLQA. For examples, both LTI and PIRCS Group can correctly answer 16 questions in such a case. The reason is not clear. Maybe different translations happen to be synonyms or paraphrases which benefit in finding answers.

8. Conclusion

This paper described an overview of NTCIR-6 CLQA. In the Formal Run, 12 groups world-wide participated in CLQA and submitted 91 runs in total. Evaluation results showed that the performance of CLQA systems were heavily degraded compared to monolingual QA systems.

However, CLQA is a new research area and low performance implies that there is a lot of room to improve the performance. It is necessary to continue NTCIR CLQA Tasks to expand the CLQA test collection as a common infrastructure and as a test bed for researchers in Cross-Lingual QA.

Acknowledgment

We thank all the participants to the NTCIR-6 CLQA task.

Part of the research project carried out in Taiwan is partially supported by Excellent Search Projects of National Taiwan University under grant no. 95R0062-AE00-02 and the Teaching Excellence Project in National Taiwan Ocean University.

Bibliography

[1] Bernardo Magnini, Simone Romagnoli, Alessandro Vallin, Jesus Herrera, Anselmo Penas, Victor Peinado, Felisa Verdejo, and Maarten de Rijke, The Multiple Language Question Answering Track at CLEF 2003, Working Notes for the CLEF 2003 Workshop, 2003.

[2] Chuan-Jie Lin, A Study on Chinese Open-Domain Question Answering System, Ph.D. dissertation, National Taiwan University, 2004.

[3] Yutaka Sasaki, Hsin-Hsi Chen, Kuang-hua Chen, Chuan-Jie Lin, Overview of the NTCIR-5 Cross-Lingual Question Answering Task (CLQA), In

Proc of NTCIR-5 Workshop Meeting, Tokyo, 2005.

[4] Satoshi Sekine and Yoshio Eriguchi, Japanese named entity extraction evaluation — analysis of results —,in Proc. of 18th International Conference on Computational Linguistics, pp. 1106-1110, 2000.

[5] Ellen M. Voorhees, The TREC-8 Question Answering Track Report, Proc. of Eighth Text REtrieval Conference (TREC-8), pp. 77-82, 1999.

Table 8. Technique descriptions of runs in Japanese-related subtasks

RunID	Index Unit	Index Tech	Index Stru	Query Unit	IR Model	Ranking	Query Expan	Trans Tech
Forst-E-J-01	Word	Morphological Analysis	Inverted file	Word	Vector space	tf/idf	No	Machine Translation, Translation dictionary, Phrase translation based on phonological information obtained from Web pages (but it is not only transliteration), Phrase translation obtained from strings that frequently appeared in titles and snippets of Web
Forst-E-J-02	Word	Morphological Analysis	Inverted file	Word	Vector space	tf/idf	No	Machine Translation, Translation dictionary, Phrase translation obtained from strings that frequently appeared in titles and snippets of Web pages, and Phrase translation based on the Wikipedia.
Forst-E-J-03	Word	Morphological Analysis	Inverted file	Word	Vector space	tf/idf	No	Machine Translation, Translation dictionary, and Phrase translation based on phonological information obtained from Web pages (but it is not only transliteration).
Forst-J-J-01	Word	Morphological Analysis	Inverted file	Word	Vector space	tf/idf	No	N/A
HARAD-J-J-01	bi-word	Lucene	Lucene	semantic graph	graph space modell	graph similarity	semantic graph	N/A
LTI-E-J-01	Word, annotations	annotations -named entities, predicate-argument	Inverted file	Morpheme, Phrase, window operator, synonym operator, weight operator	Language Modeling and inference network (Indri)	KL/cross entropy (Indri)	Multiple translation candidates, predicate-argument expansion	Question sentence translation using webMT, multiple translation candidates for named entities terms using wikipedia, webMT's, and web-mining
LTI-E-J-02	Word, annotations	annotations -named entities, predicate-argument	Inverted file	Morpheme, Phrase, window operator, synonym operator, weight operator	Language Modeling and inference network (Indri)	KL/cross entropy (Indri)	Multiple translation candidates, predicate-argument expansion	Question sentence translation using webMT, multiple translation candidates for named entities terms using wikipedia, webMT's, and web-mining
LTI-E-J-03	Word, annotations	annotations -named entities, predicate-argument	Inverted file	Morpheme, Phrase, window operator, synonym operator, weight operator	Language Modeling and inference network (Indri)	KL/cross entropy (Indri)	Multiple translation candidates, predicate-argument expansion	Question sentence translation using webMT, multiple translation candidates for named entities terms using wikipedia, webMT's, and web-mining
LTI-J-J-u-01	Word, annotations	annotations -named	Inverted file	Morpheme, Phrase, window	Language Modeling and	KL/cross entropy	Keyword expansion,	N/A

		entities, predicate-argument		operator, synonym operator, weight operator	inference network (Indri)Block (3 sentence chunk) retrieval	(Indri)	Predicate-argument expansion	
LTI-J-J-u-02	Word, annotations	annotations -named entities, predicate-argument	Inverted file	Morpheme, Phrase, window operator, synonym operator, weight operator	Language Modeling and inference network (Indri)Block (3 sentence chunk) retrieval	KL/cross entropy (Indri)	Keyword expansion, Predicate-argument expansion	N/A
TITFL-E-J-01	character+bi-character+tri-character	character segmentation	inverted file	character+bi-character+tri-character	vector space model	tf/idf	no	web translation tool
TITFL-E-J-02	character+bi-character+tri-character	character segmentation	inverted file	character+bi-character+tri-character	boolean model+vector space model	tf/idf	no	web translation tool
TITFL-E-J-03	character+bi-character+tri-character	character segmentation	inverted file	character+bi-character+tri-character	boolean model	tf/idf	no	web translation tool
TITFL-J-J-01	character+bi-character+tri-character	character segmentation	inverted file	character+bi-character+tri-character	vector space model	tf/idf	no	n/a
TITFL-J-J-02	character+bi-character+tri-character	character segmentation	inverted file	character+bi-character+tri-character	boolean model+vector space model	tf/idf	no	n/a
TTH-E-J-01	English word	stemming	inverted file	English word	vector space model	SMART	no	statistical MT based
TTH-E-J-02	English word	stemming	inverted file	English word	vector space model	SMART	no	statistical MT based
TTH-E-J-03	Japanese word, bi-word, bi-char	morphology	inverted file	Japanese word, bi-word, bi-char	OKAPI BM25	TF.IDF	pseudo relevance feedback	transliteration and dictionary based
TTH-J-J-01	Japanese word, bi-word	morphology	inverted file	Japanese word	vector space model	SMART	no	no
TTH-J-J-02	Japanese and English word	morphology, stemming	inverted file	Japanese and English word	vector space model	SMART	English word translation	MT based
TTH-J-E-01	Japanese word	morphology	inverted file	Japanese word	vector space model	SMART	no	statistical MT based

Table 9. Technique descriptions of runs in Chinese-related subtasks

RunID	Index Unit	Index Tech	Index Stru	Query Unit	IR Model	Ranking	Query Expan	Trans Tech
IASL-C-C-01	Word and Char	N/A	Inverted File	Word and Char	Vector space	tf/idf	No	N/A
IASL-C-C-02	Word and Char	N/A	Inverted File	Word and Char	Vector space	tf/idf	No	N/A
IASL-C-C-03	Word and Char	N/A	Inverted File	Word and Char	Vector space	tf/idf	No	N/A
IASL-E-C-01	Word and Char	N/A	Inverted File	Word and Char	Vector space	tf/idf	No	Google Translate
IASL-E-C-02	Word and Char	N/A	Inverted File	Word and Char	Vector space	tf/idf	No	Google Translate
IASL-E-C-03	Word and Char	N/A	Inverted File	Word and Char	Vector space	tf/idf	No	Google Translate
ICDCU-C-C-01	word	Stopword +dictionary	Inverted file	Word	Probabilistic	BM25	No	QE
NTCIR6-ILS-Run1	word	N/A	N/A	N/A	Lucene	N/A	N/A	AltaVista Babel Fish
MHC	word	N/A	N/A	Structured query operators	Hanquery	N/A	N/A	Dictionary, bi-words
LTI-C-	Word,	Word -	Inverted	Word	Language	KL/cross-	N/A	N/A

C-01	annotations	segmented, annotations - named entities	file		Modeling and inference network (Indri)	entropy(Indri) + maxent based discriminative reranking		
LTI-C-C-02	Word, annotations	Word - segmented, annotations - named entities	Inverted file	Word	Language Modeling and inference network (Indri)	KL/cross-entropy(Indri) + maxent based discriminative reranking	Google expansion on keywords	N/A
LTI-C-C-03	Word, annotations	Word - segmented, annotations - named entities	Inverted file	Word	Language Modeling and inference network (Indri)	KL/cross-entropy(Indri) + pair-wise ranking SVM		N/A
LTI-E-C-01	Word, annotations	Word - segmented, annotations - named entities	Inverted file	Word	Language Modeling and inference network (Indri)	KL/cross-entropy(Indri) + maxent based discriminative reranking	Multiple keyword translation candidates + multiple full sentence translation	Question sentence translation using multiple webMT, keyterm translation using dictionaries, wikipedia, webMT's, and web-mining
LTI-E-C-02	Word, annotations	Word - segmented, annotations - named entities	Inverted file	Word	Language Modeling and inference network (Indri)	KL/cross-entropy(Indri) + maxent based discriminative reranking	Multiple keyword translation candidates + multiple full sentence translation +Google expansion	Question sentence translation using multiple webMT, keyterm translation using dictionaries, wikipedia, webMT's, and web-mining
LTI-E-C-03	Word, annotations	Word - segmented, annotations - named entities	Inverted file	Word	Language Modeling and inference network (Indri)	KL/cross-entropy(Indri) + pair-wise ranking SVM	Multiple translation candidates	Question sentence translation using multiple webMT, keyterm translation using dictionaries, wikipedia, webMT's, and web-mining
NCUTW-C-C-01	bi-character	Stopword removal	inverted file	bi-character	probabilistic model (BM-25)	Okapi BM-25	N/A	N/A
NCUTW-E-C-01	bi-character	Stopword removal	inverted file	bi-character	probabilistic model (BM-25)	Okapi BM-25	N/A	Question Translation with Systran online
pircs-C-C-01	1gram + bigram	5stopchar	Inverted File	1gram + bigram	PIRCS	tf.ICTF	NoQryExpan	n.a.
pircs-C-C-02	1gram + bigram	5stopchar	Inverted File	1gram + bigram	PIRCS	tf.ICTF	NoQryExpan	n.a.
pircs-C-C-03	1gram + bigram	5stopchar	Inverted File	1gram + bigram	PIRCS	tf.ICTF	NoQryExpan	n.a.
pircs-E-C-04	1gram + bigram	5stopchar	Inverted File	1gram + bigram	PIRCS	tf.ICTF	NoQryExpan	SystranMT + Web-based entity Trans
pircs-E-C-05	1gram + bigram	5stopchar	Inverted File	1gram + bigram	PIRCS	tf.ICTF	NoQryExpan	SystranMT + Web-based entity Trans
pircs-E-C-06	1gram + bigram	5stopchar	Inverted File	1gram + bigram	PIRCS	tf.ICTF	NoQryExpan	SystranMT + Web-based entity Trans
WMMKS-C-C-01	bi-word (Chinese Corpus)	POS	inverted file	phrase	vector space	TF, phrase length, wikipedia article, Google Search Results	none	none
WMMKS-C-E-01	word (English corpus)	POS	inverted file	phrase	vector space	TF, phrase length, wikipedia article, Google Search Results	none	Google Translation
WMMKS-E-C-01	bi-word (Chinese Corpus)	POS	inverted file	phrase	vector space	TF, phrase length, wikipedia article, Google Search Results	none	Google Translation