Overview of the NTCIR-5 Cross-Lingual Question Answering Task (CLQA1)

Yutaka Sasaki
ATR Spoken Language Communication Research Laboratories
2-2-2 Hikaridai, Keihanna Science City, Kyoto, 619-0288 Japan
yutaka.sasaki@atr.jp

Hsin-Hsi Chen, Kuang-hua Chen
National Taiwan University
No 1, Roosevelt Rd. Sec 4, Taipei 10617, Taiwan
hh_chen@csie.ntu.edu.tw, khchen@ntu.edu.tw

Chuan-Jie Lin
National Taiwan Ocean University
No 2, Pei-Ning Rd., Keelung 20224, Taiwan
cjlin@ntou.edu.tw

Abstract

This paper gives an overview of the NTCIR-5 Cross-Lingual Question Answering Task (CLQA1), an evaluation campaign for Cross-Lingual Question Answering technology. This evaluation was carried out in June 2005. In CLQA1, we aimed to promote research on cross-lingual Question Answering technology mainly for East Asian languages.

As the first attempt, we conducted evaluations of five subtasks: JE, EJ, CE, CC, and EC subtasks, where C, E, and J stand for Chinese, English, and Japanese, respectively, and XY 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 200-300 sample question/answer pairs for each subtask. The Formal Run evaluation was conducted during June 13-27, 2005 with 200 test questions. As a result, 13 research institutes world-wide participated in CLQA1, and 89 runs were submitted in total.

Keywords: *CLQA, Question Answering, NTCIR*

1 Introduction

Question Answering has recently been actively studied all over the world since the start of the Question Answering Track at TREC-8 [4]. In 2003, CLEF started the Multilingual Question Answering Track for European languages, such as Italian, Spanish, German, Dutch, and French, as QA@CLEF. [1]

In a series of NTCIR Workshops, the CLIR and QAC Tasks have been conducted for three years. So, it is now time to step forward and attempt to promote technologies of *Cross-Lingual Question Answering (CLQA)* for *Asian languages* based on a common test set, in addition to the CLIR and QAC Tasks as continuing tasks. In light of this, NTCIR-5 (2004/2005) initiated the NTCIR CLQA Task as a pilot task. From a linguistic viewpoint, CLQA is a much more complicated challenge. As a result, we decided to make the question target much simpler; Only questions about Named Entities are provided in the NTCIR CLQA1 Task.

2 Overview

As the first attempt in the NTCIR CLQA1 Task, we conducted an evaluation of five subtasks: JE, EJ, CE, CC, and EC subtasks, where C, E, and J stand for Chinese, English, and Japanese, respectively, and XY indicates that questions are given in language X (*source language*) and answers must be extracted from documents written in language Y (*target language*). Note that an evaluation corresponding to the JJ subtask was separately conducted in NTCIR QAC.

Target Documents CLQA1 provided participants with three kinds of corpora for the three languages.

1. Chinese: news articles spanning a period of two years (2000 and 2001) taken from UDN.COM. (A total of 901,446 news articles)

Table 1. Corpora Used in Each Subtask

Source \ Target	C	E	J
C	UDN	Daily Yomiuri	—
E	UDN	—	Yomiuri Newspaper
J	—	Daily Yomiuri	—

2. Japanese: news articles spanning a period of two years (2000 and 2001) taken from the Yomiuri Newspaper. (A total of 658,719 news articles.[1])

3. English: news articles spanning a period of two years (2000 and 2001) taken from the Daily Yomiuri. (A total of 17,741 news articles.[2])

Table 1 shows the corpora used for each subtask.

Scope of Answer 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 [3] in 1999.[3]

The NEs defined by IREX are:

1. PERSON
2. LOCATION
3. ORGANIZATION
4. ARTIFACT (product name, book title, law, ...)
5. DATE
6. TIME
7. MONEY
8. PERCENT

We adopt these NEs plus NUMEX for CLQA1:

9. 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 Questions are created according to subtasks as follows.

JE and EJ subtasks: Since the Daily Yomiuri articles are English translations of Yomiuri Shimbun articles, we first manually selected corresponding articles between the two corpora[4] and 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.

CE subtask: Chinese questions of the CE subtask are Chinese translations of the English questions for the EJ subtask. This is because CLQA1 employed the Daily Yomiuri as an English knowledge source.

CC and EC 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 [5] with filtering out non-NE questions, and questions which seemed not to have answers in the UDN collection.[6] And then, Chinese questions were translated into English for EC subtask.

We provided as sample data 300 question/answer pairs for the EJ, JE, and CE subtasks and 200 pairs for the CC and EC subtasks.

For the Formal Run evaluation, we provided 200 questions for each subtask. Table 2 shows the number of questions for each question type.

Schedule The time schedule of CLQA1 was as follows:

[1] Two broken articles and 1,701 empty articles were removed from the original article database of 2000 and 1,870 empty articles were removed from the original article database of 2001.

[2] One empty article was removed from the original article database of 2000.

[3] http://nlp.cs.nyu.edu/irex/index-e.html

[4] There is no link between the two corpora.

[5] http://nlg.csie.ntu.edu.tw/ [2]

[6] It was decided by roughly searching UDN articles by question creators.

Table 2. Question Type Distribution of Formal Run Questions

Category \ Subtask	JE/EJ/CE	CC/EC
PERSON	27	80
LOCATION	30	53
ORGANIZATION	26	18
ARTIFACT	18	13
DATE	25	0
TIME	14	20
MONEY	20	4
PERCENT	10	0
NUMEX	30	12
Total	200	200

1. 2004-09-14: Round table discussion about CLQA
2. 2004-11-01: Call for participations delivered
3. 2004-11-30: Deadline for application
4. 2004-12-15: Release of document Sets
5. 2005-02-04: Release of Q/A sample data for JE/EJ subtasks
6. 2005-02-19: Release of Q/A sample data for CE/CC/EC subtasks
7. 2005-06-12: System development freeze
8. 2005-06-13/20: Formal run evaluation period
9. 2005-09-03: Delivery of manual evaluation results for JE/EJ subtasks
10. 2005-09-11: Delivery of manual evaluation results for CE/CC/EC subtasks
11. 2005-10-15: Deadline for workshop proceedings
12. 2005-12-06/09: NTCIR Workshop 5

3 Participants

In total, 13 groups participated in CLQA1, 9 for Chinese related subtasks and 7 for Japanese related subtasks. Table 3 shows the number of formal runs submitted by participants. The asterisk (*) indicates subtasks from which participants withdrew.

We accepted submissions of at most three official runs and unlimited unofficial runs for each subtask. The parentheses show the number of submitted unofficial runs.

4 Task Definitions

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 CLQA1, the character encoding of the input was BIG5 for Chinese, US-ASCII for English, and EUC-JP for Japanese. The input format of CLQA1 is defined as follows.

```
[QID]: "[Question]"
```

QID is the form of [QuestionSetID]-[Lang]-[QuestionNo]-[SubQuestionNo].

QuestionSetID is ”CLQA1”.

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:
  CLQA1-EN-S0001-00: "When did Queen
  Victoria die?"
```

The character encoding of the output was BIG5 for Chinese, US-ASCII for English, and EUC-JP for Japanese. CLQA1 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)

Table 3. Submitted Runs

Group ID	Region	JE	EJ	CE	CC	EC	Total
DCU (LCC-DCU)	Europe			*	*	*	
DLTG	Europe				1(0)		1(0)
Forst	Asia	3(4)	3(4)				6(8)
IASL	Asia				3(5)		3(5)
ICT (LCC-DCU)	Asia			*	1(1)	*	1(1)
ISCAS	Asia			*	*	*	
LTI	US		3(3)			3(3)	6(6)
NCQAL	Asia	1(1)	*				1(1)
NICT	Asia	0(2)	1(2)				1(4)
ntoua	Asia			3(3)			3(3)
NYU	US	*	*				
pircs	Asia					3(3)	3(3)
QATRO	Asia	3(3)	3(3)	3(4)	*	*	9(10)
SYIAE	Asia			*	*	*	
TTN	Asia	1(1)	2(2)				3(3)
UNTIR	US			1(1)	1(1)	1(1)	3(3)
WMMKS	Asia			*	1(0)	1(0)	2(0)
Total		8(11)	12(14)	7(8)	7(7)	8(7)	42(47)
# Groups		5	5	3	5	4	13

[official runs (unofficial runs)]

```
CLQA1-EN-S0001-00, EN, "1901",
ENY-20001101CYM0398, ,
CLQA1-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 golden files, is encoded in UTF-8. The format of the test collection for CLQA1 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 CLQA1.

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 CLQA1. 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 if no answer.

Example:

```
<?xml version="1.0" encoding="UTF-8"?>
<QASET>
<VERSION>NTCIR-5 CLQA1 Training Set
v1.0-R1.0 (2005.2.1)</VERSION>
<QA>
<QUESTION>
<Q LANG="EN" QID="CLQA1-EN-S0001-00">
When did Queen Victoria die?</Q>
<Q LANG="JA" QID="CLQA1-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, as it is the first attempt this year, 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. However, we still encouraged participants to submit translated answers in unofficial runs in order to learn the possibility of running an answer-translated task.

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:

```
CLQA1-EN-T0001-00, EN, "1901",
ENY-20001101CYM0398, ,
CLQA1-EN-T0002-00, EN
CLQA1-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:

```
CLQA1-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 intend to find answers of English questions in Chinese documents, the output for the same question looks like:

```
CLQA1-EN-T0003-00, ZH, "張三",
mhn_xxx_20010808_1034915, ,
CLQA1-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.

1. RunID: as explained in the RunID Section.
2. IndexUnit: character, bi-character, bi-word, phrase, etc.
3. IndexTech: the techniques used to process index terms, e.g., morphology, stemming, POS, etc.
4. IndexStruc: inverted file, signature file, PAT, etc.
5. QueryUnit: character, word, phrase, etc.
6. IRModel: vector space model, probabilistic model, etc.
7. Ranking: ranking factor for measuring each term, e.g., tf, tf/idf, mutual information, word association, document length, etc.
8. QueryExpan: techniques used to expand query or no query expansion
9. 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:

1. Right (R): the answer is correct, and the document where it is from supports it.
2. 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.
3. 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 pool of runs of JE/EJ subtasks was conducted independent of the organizers by a Japanese company specializing foreign language communication.

The assessment of Chinese related subtasks was conducted as follows. For CC and EC subtasks, two assessors were asked to judge all [answer, docID] pairs proposed in all of the runs, including official and unofficial ones. If there was an inconsistent judgment among the pool, between the two assessors, or with the previous prepared answer, a third assessor (which was Chuan-Jie Lin himself this year) would do the final judgment. For CE subtask, due to the time and effort limit, only one assessor was involved in the assessment. Again, if the judgment was inconsistent, a second assessor 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 is the rate at which the 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 JE/EJ Subtasks

Tables 5-6 show the evaluation results of JE/EJ subtasks. The asterisk (*) indicates runs submitted by one of CLQA1 organizers.

In JE subtask, 8 official runs and 11 unofficial runs were submitted from 5 institutes. The best official run was submitted by NCQAL group. The accuracy was 30.0% with counting only supported answers. It rises to 31.5% if unsupported answers were considered correct.

In EJ subtask, 12 official runs and 14 unofficial runs were submitted from 5 institutes. The best official run was submitted by Forst group. The accuracy was 12.5% with counting only supported answers. It rises to 15.5% if unsupported answers were considered correct. Due to the limitation of evaluation resources, unofficial runs of JE/EJ subtasks were not able to be evaluated.

6.2 Results of CE/CC/EC Subtasks

Tables 7-12 show evaluation results of the CE/CC/EC subtasks. The asterisk indicates runs submitted by one of CLQA1 organizers.

In CC subtask, 7 official runs and 7 unofficial runs were submitted from 5 institutes. The best three official runs were submitted by IASL group, which accuracies were 33% to 37.5%. The second best group was WMMKS, which had a similar performance as the top group. Other groups achieved 10% to 14% accuracy.

In EC subtask, 8 official runs and 7 unofficial runs were submitted from 4 institutes. The best three official runs were submitted by PIRCS group, which accuracies were 11.5% to 12.5%. The second best three official runs were submitted by LTI group, which accuracies were 5% to 7.5%.

In CE subtask, 7 official runs and 8 unofficial runs were submitted from 3 institutes. The performances in CE subtask were somewhat lower than those in EC subtask. The best accuracy score was only 6% by UNTIR group, and the others' scores ranged from 1% to 3%.

6.3 Analysis Results

Figures 1 and 2 show the ratio of correct answers for each question.

7 Discussion

CLQA1 has started to extend the QA framework to a Cross-Lingual QA framework. From the viewpoint of a research area, CLQA is the research direction that merges Machine Translation research and Question Answering research.

From our experience in organizing CLQA1, CLQA is a more challenging target than monolingual QA. Because of the translation phase, there are several difficulties in CLQA systems.

1. Translated questions are represented with different expressions than those used in news articles in which answers appear.

Table 4. Monolingual vs. Cross-lingual

Run ID	# Correct		Accuracy	
	# R	# R+U	R	R+U
Forst-E-J-03	25	31	12.5	15.5
Forst-J-J	34	53	17.0	26.5
			+4.5	+11.0
LTI-E-J-02	20	25	10.0	12.5
LTI-J-J	16	40	8.0	20.0
			-2.0	+7.5
NCQAL-J-E-01	60	63	30.0	31.5
NCQAL-E-E	74	85	37.0	42.5
			+7.0	+11.0
NICT-E-J-01	18	24	9.0	12.0
NICT-J-J	34	53	17.0	26.5
			+8.0	+14.5
QATRO-E-J-01*	0	1	0.0	0.5
QATRO-J-J	4	9	2.0	4.5
			+2.0	+4.0
QATRO-J-E-01*	2	2	1.0	1.0
QATRO-E-E	11	16	5.5	8.0
			+4.5	+7.0
TTN-E-J-01	11	13	5.5	6.5
TTN-J-J	22	35	11.0	17.5
			+5.5	+11.0

2. Since key words for retrieving documents are translated words from an original question, document retrieval in CLQA becomes much more difficult than that in monolingual QA.

In JE/EJ subtasks, Forst and TTN teams used Japanese QA systems for both JE and EJ CLQA subtasks. This was a successful approach for EJ subtask but not for JE subtasks. Another interesting point is how much CLQA systems degrade from monolingual QA systems. Table 4 shows a comparison between Japanese and English monolingual QA systems, i.e., equivalent to JJ/EE subtasks, and CLQA systems. This table is created based on reports from the participants of JE/EJ subtasks.

As a result, it is revealed that around 5-10% degradation was caused by translation of questions or documents.

In CC subtask, some groups mentioned that it was their first time to develop Chinese QA systems (or NLP systems in traditional Chinese). They used simple rules to do question classification and answer extraction, and employed systems developed by others to process Chinese texts. On the other hand, the best two groups, IASL and WMMKS, had more experiences in Chinese NLP and QA. It seems that the performance of monolingual QA depends on the achievements in NLP technologies.

Considering the translation techniques used in the EC and CE subtasks, machine-readable dictionaries and online MT systems were commonly used. Some groups also used search results from the web to do translation. Co-occurrences of translation pairs were adopted but in different ways. PIRCS also used the web to do query expansion and considered a frequent name in the search results to be a possible answer.

Comparing monolingual and cross-lingual QA, the performances in EC subtask are approximately only 1/3 of the performances in CC subtask. It shows certain challenges in CLQA.

There were only two groups participated both in CC and EC subtasks this year. The WMMKS team achieved 32% and 8% accuracies in CC and EC subtasks, respectively. But they found some bugs in their system. Hence the results might not stand for their real abilities in monolingual and cross-lingual QA. The UNTIR group participated in all CC, EC, and CE subtasks, and achieved 10%, 3%, and 6% accuracies, respectively. But they mentioned that their systems were not fully developed, considering the poor performance in CC subtask. It was a pity that no comparable results could derive any conclusion this year.

We found that there were more challenges in CE subtask than in EC subtask. Because the Chinese questions in CE subtask were created by translating the English questions in EJ subtask, which had corresponding Japanese questions in JE subtask. The questions in EJ/JE subtasks were created by reading the news articles from Yomiuri, a Japanese news agent. Many of the questions asked things happening in Japan, and many of the named entities in the Chinese questions were originally Japanese names which might or might not use the same Kanji (or Chinese characters).

In order to achieve good performance, a system should have ability to identify both Chinese and Japanese names in a Chinese question, and find its English translation not only by using Chinese resources but also by Japanese resources. For example, in question CLQA1-ZH-T0009-00, the person name “中澤” (Nakazawa) uses different Kanji as in Japanese, “中沢”. Besides, “中澤” is not a Chinese name, hence a Chinese NE system may fail to identify it as a person name.

Another example in the same question is the organization name “市立船橋高中” (Ichiritsu Funabashi High School). It refers to a high school in Japan. Without Japanese resources, the term “市立” will be translated into “municipal”, and “船橋” will be translated into “Boat-Bridge” (by meaning) or “Chuanqiao” (by sound). Some groups used web search engines to find corresponding translations. If a name was written differently in Japanese and Chinese, they had to recover its original name then did searching in Japanese web pages, since the name might not appear in a Chinese web page.

In order to focus on the development of QA techniques, we will try to minimize such NE problems in the future. However, in order to process texts written in CJK characters, developing a CJK NE identifier or

techniques to find corresponding CJK translations is necessary.

8 Conclusion

This paper described an overview of NTCIR-5 CLQA1. In the Formal Run, 13 groups world-wide participated in CLQA1 and submitted 89 runs in total. Evaluation results showed that the performance of CLQA systems were 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.

References

[1] Bernardo Magnini, Simone Romagnoli, Alessandro Vallin, Jesús Herrera, Anselmo Peñas, Víctor 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] 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.

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

Table 5. JE Subtask

Run ID	# Correct		Accuracy	
	# R	# R+U	R	R+U
Forst-J-E-01	6	6	3.0	3.0
Forst-J-E-02	17	18	8.5	9.0
Forst-J-E-03	16	17	8.0	8.5
NCQAL-J-E-01	60	63	30.0	31.5
QATRO-J-E-01*	2	2	1.0	1.0
QATRO-J-E-02*	2	2	1.0	1.0
QATRO-J-E-03*	1	1	0.5	0.5
TTN-J-E-01	0	0	0.0	0.0

Table 6. EJ Subtask

Run ID	# Correct		Accuracy	
	# R	# R+U	R	R+U
LTI-E-J-01	18	21	9.0	10.5
LTI-E-J-02	20	25	10.0	12.5
LTI-E-J-03	17	20	8.5	10.0
Forst-E-J-01	13	18	6.5	9.0
Forst-E-J-02	18	28	9.0	14.0
Forst-E-J-03	25	31	12.5	15.5
NICT-E-J-01	18	24	9.0	12.0
QATRO-E-J-01*	0	1	0.0	0.5
QATRO-E-J-02*	0	0	0.0	0.0
QATRO-E-J-03*	0	0	0.0	0.0
TTN-E-J-01	11	13	5.5	6.5
TTN-E-J-02	11	13	5.5	6.5

Table 7. CE Subtask

Run ID	# Correct		Accuracy	
	# R	# R+U	R	R+U
nouta-C-E-01*	3	4	1.5	2.0
nouta-C-E-02*	5	6	2.5	3.0
nouta-C-E-03*	6	7	3.0	3.5
QATRO-C-E-01*	5	5	2.5	2.5
QATRO-C-E-02*	3	3	1.5	1.5
QATRO-C-E-03*	2	2	1.0	1.0
UNTIR-C-E-01	12	13	6.0	6.5

Table 8. CC Subtask

Run ID	# Correct		Accuracy	
	# R	# R+U	R	R+U
DLTG-C-C-01	28	30	14.0	15.0
IASL-C-C-01	75	89	37.5	44.5
IASL-C-C-02	68	74	34.0	37.0
IASL-C-C-03	66	72	33.0	36.0
lcc-C-C-01	20	21	10.0	10.5
UNTIR-C-C-01	20	21	10.0	10.5
WMMKS-C-C-01	64	70	32.0	35.0

Table 9. EC Subtask

Run ID	# Correct		Accuracy	
	# R	# R+U	R	R+U
LTI-E-C-01	14	19	7.0	9.5
LTI-E-C-02	15	19	7.5	9.5
LTI-E-C-03	10	12	5.0	6.0
pirc-E-C-01	25	33	12.5	16.5
pirc-E-C-02	23	28	11.5	14.0
pirc-E-C-03	24	30	12.0	15.0
UNTIR-E-C-01	6	8	3.0	4.0
WMMKS-E-C-01	8	9	4.0	4.5

Table 10. CE Subtask (Unofficial Results)

	# R					# R+U					R (%)			R+U (%)		
Run	1	2	3	4	5	1	2	3	4	5	Acc	MRR	Top5	Acc	MRR	Top5
ntoua-C-E-u-01*	5	6	3	3	4	6	6	3	3	4	2.5	5.3	10.5	3.0	5.8	11.0
ntoua-C-E-u-02*	6	3	1	1	2	7	4	2	1	3	3.0	4.2	6.5	3.5	5.3	8.5
ntoua-C-E-u-03*	7	4	5	1	1	8	4	7	2	1	3.5	5.6	9.0	4.0	6.5	11.0
QATRO-C-E-u-01*	5	6	4	3	2	5	7	4	3	2	2.5	5.2	10.0	2.5	5.5	10.5
QATRO-C-E-u-02*	3	4	6	3	2	3	4	6	3	2	1.5	4.1	9.0	1.5	4.1	9.0
QATRO-C-E-u-03*	2	5	1	2	3	2	5	1	2	3	1.0	3.0	6.5	1.0	3.0	6.5
QATRO-C-E-u-04*	2	1	1	5	2	2	1	1	5	2	1.0	22.4	5.5	1.0	2.2	5.5
UNTIR-C-E-u-01	12	7	7	1	3	13	8	7	2	3	6.0	9.3	15.0	6.5	10.2	16.5

Table 11. CC Subtask (Unofficial Results)

	# R					# R+U					R (%)			R+U (%)		
Run	1	2	3	4	5	1	2	3	4	5	Acc	MRR	Top5	Acc	MRR	Top5
IASL-C-C-u-01	48	0	0	0	0	51	0	0	0	0	24.0	24.0	24.0	25.5	25.5	25.5
IASL-C-C-u-02	63	0	0	0	0	67	0	0	0	0	31.5	31.5	31.5	33.5	33.5	33.5
IASL-C-C-u-03	40	0	0	0	0	44	0	0	0	0	20.0	20.0	20.0	22.0	22.0	22.0
IASL-C-C-u-04	58	0	0	0	0	63	0	0	0	0	29.0	29.0	29.0	31.5	31.5	31.5
IASL-C-C-u-05	60	0	0	0	0	64	0	0	0	0	30.0	30.0	30.0	32.0	32.0	32.0
LCC-C-C-u-02	47	0	0	0	0	51	0	0	0	0	23.5	23.5	23.5	25.5	25.5	25.5
UNTIR-C-C-u-01	20	6	4	1	1	21	7	4	2	1	10.0	12.4	16.0	10.5	13.3	17.5

Table 12. EC Subtask (Unofficial Results)

	# R					# R+U					R (%)			R+U (%)		
Run	1	2	3	4	5	1	2	3	4	5	Acc	MRR	Top5	Acc	MRR	Top5
LTI-E-C-u-01	14	11	6	5	5	19	13	10	5	5	7.0	11.9	20.5	9.5	15.5	26.0
LTI-E-C-u-02	15	11	6	4	5	19	15	9	5	5	7.5	12.3	20.5	9.5	15.9	26.5
LTI-E-C-u-03	10	10	13	3	6	12	15	16	4	7	5.0	10.6	21.0	6.0	13.6	27.0
pircs-E-C-u-04	25	14	14	6	6	33	20	15	8	7	12.5	19.7	32.5	16.5	25.7	41.5
pircs-E-C-u-05	23	15	14	6	8	28	18	15	10	9	11.5	19.1	33.0	14.0	23.1	40.0
pircs-E-C-u-06	24	13	12	4	9	30	16	14	5	10	12.0	18.7	31.0	15.0	23.0	37.5
UNTIR-E-C-u-01	6	3	2	0	0	8	3	3	1	0	3.0	4.1	5.5	4.0	5.4	7.5

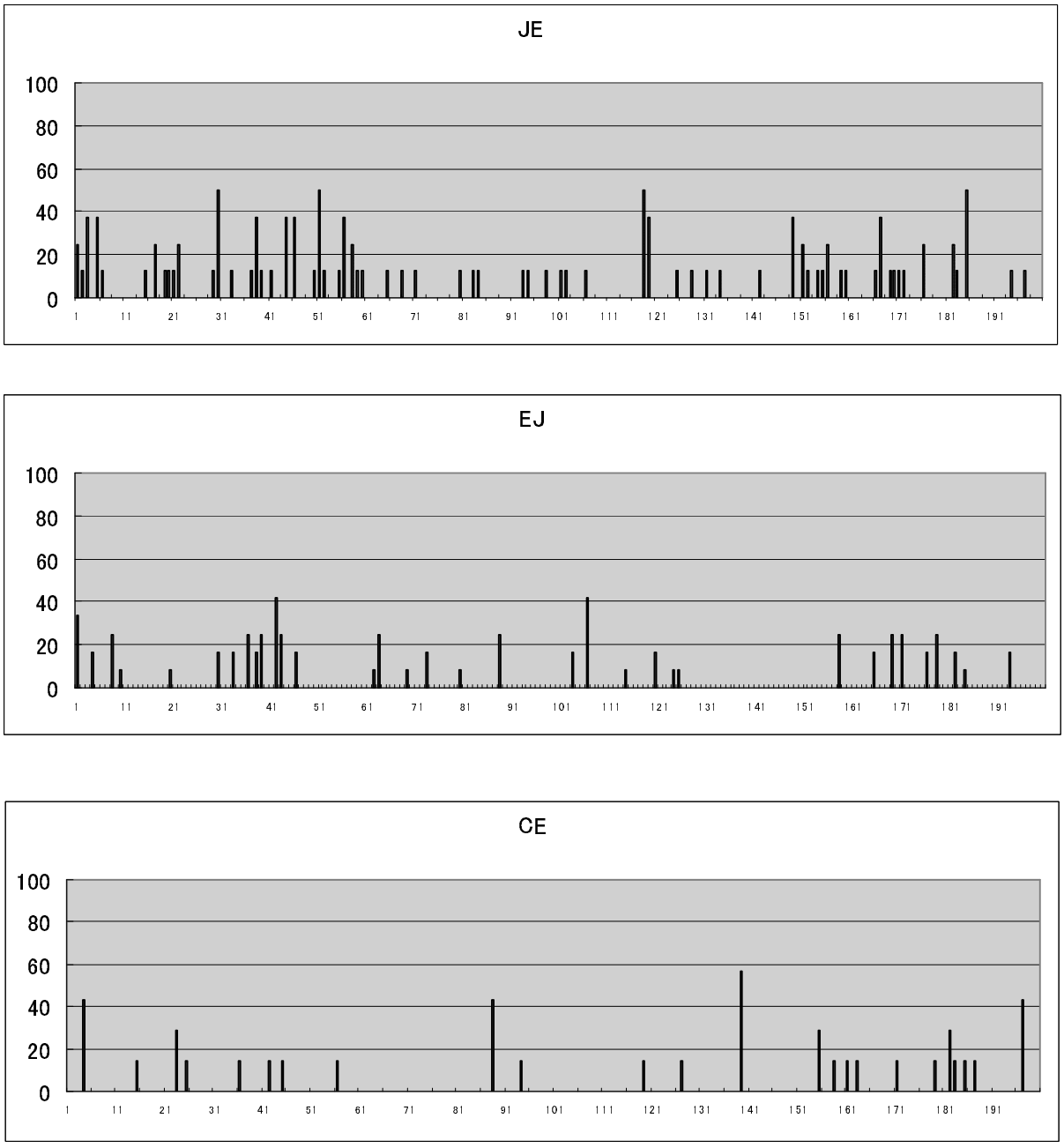

Figure 1. Ratio (%) of Correct Answers for Each Question

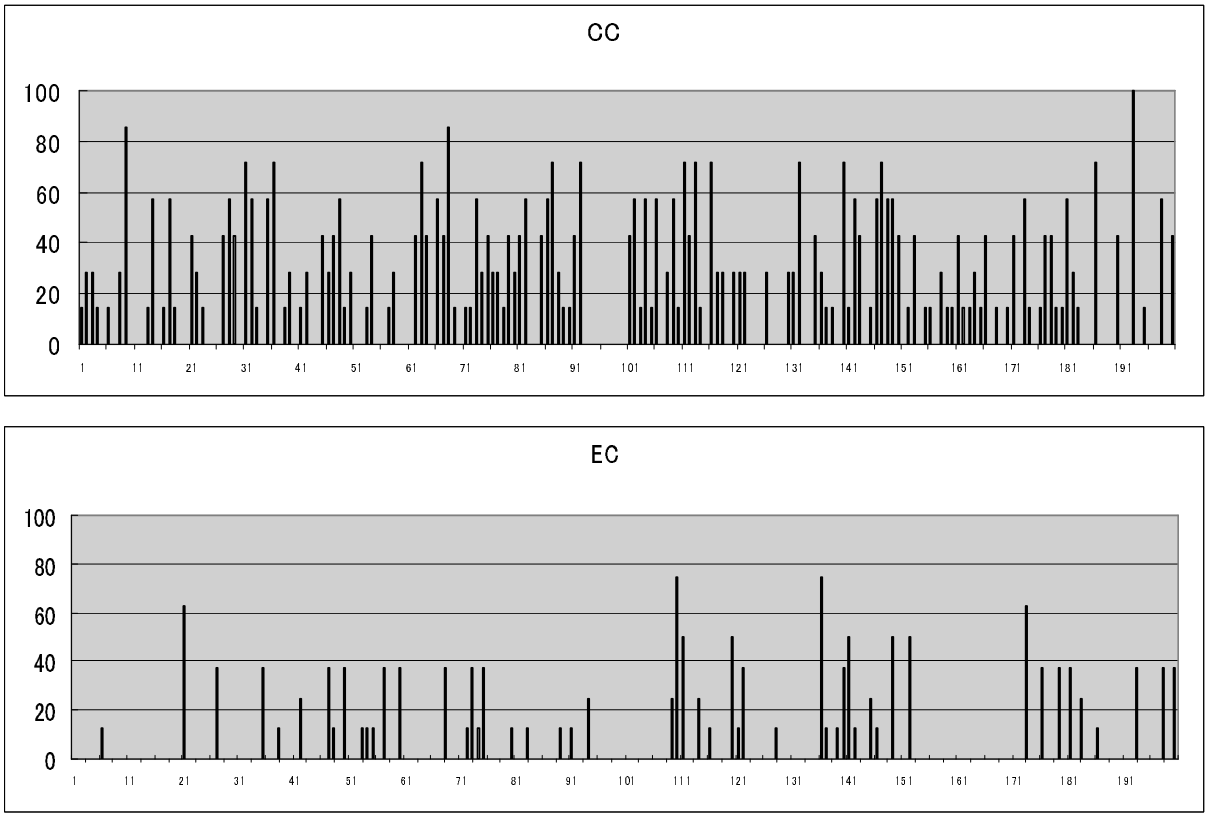

Figure 2. Ratio (%) of Correct Answers for Each Question (cont.)