NYU/CRL QA system, QAC question analysis and CRL QA data

Satoshi SEKINE   Kiyoshi SUDO   Yusuke SHINYAMA
New York University
715 Broadway, 7th floor, New York, NY 10003, USA
`sekine,sudo,yusuke@cs.nyu.edu`

Chikashi NOBATA   Kiyotaka UCHIMOTO   Hitoshi ISAHARA
Communications Research Laboratory
2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan
`nova,uchimoto,isahara@crl.go.jp`

August 17, 2002

Abstract

In this paper, we will describe NYU/CRL QA system which is a system to participate QAC Task-1 evaluation and analysis of the result. Then two analyses of QAC questions are reported. One is a categorization of the NE type of answers and the other is investigation of human performance on the QAC questions. Also, we will introduce the CRL QA data which we created by the similar method to QAC data and also which we made publicly available.

Keywords: QA system, QA data

1   Introduction

In this paper, we will describe the following three issues.

1. NYU/CRL QA system description
   We participated QAC Task-1 (only). The description of the system, the evaluation result and analysis of the result will be reported.

2. Analysis of QAC questions
   We analyzed the questions used in the QAC Task-1.

   First, we categorize the NE type of the answers. NE type of answers are categorized based on 140 NE categories used in our system and IREX's 8 categories [IREX Committee 1999] [Sekine and Isahara 2000]. The distribution NE types is compared to the distribution of NE in general newspaper in IREX evaluation.

   Second, we will report the human performance for QAC questions using their own knowledge only and using Web knowledge. The result may suggest the possible upper limit and bottom line performance systems may be expected to have. Also, in the process of adjusting the human answers to the correct answers, whose strings are restricted to those appeared in the target newspaper articles, we found that the variety of expressions happens quite often, but the variations are not so irregular. We think this is a good resource to the discussion of future evaluation on the usage of different sources and if answers have to be exact extractions from source data or not.

3. CRL QA data
   We built QA data, called CRL-QA data, which consists of 2000 question and (possibly more than one) answer pairs. The data also include question type, NE type and other information. It is available at `http://cs.nyu.edu/ŝekine/PROJECT/CRLQA`. The data is created almost the same manner to the QAC data. The differences are 1) it includes wider variety of questions including those without interrogative, 2) answers are not checked by multiple systems, so the coverage may not be sufficient, 3) more information are included and the number of questions is larger.

2   NYU/CRL QA system

In this section our QA system and the evaluation result are described.

The NYU/CRL QA system consists of three components.

1. Question Examination (QE) Examine the question sentence using question patterns. It creates information like keywords and NE type.

2. Text Retrieval (TR) Based on keywords, texts which is expected to have answers are retrieved.

3. Answer Extraction (AE) Among retrieved texts, answer strings will be searched using the information created by the QE component.

Note that we did NOT use the Mainichi 98 and 99 corpus in the knowledge creation etc by any means for any purpose. In the following description, 'training QA data' means QAC dev data and CRL-QA data which will be explained later. We will describe each components in the following sub sections.

2.1 Question Examination

The input of this component is the question sentence and the output is the list of keywords, NE types of the answer, and several kinds of minor information. The sentence is first analyzed by morphological analyzer, JUMAN[Juman], and our NE tagger. Words are concatenated if it is a sequence of noun-prefix, nouns and noun-suffix or a NE expression, while the individual words remains (which are used as keyword etc with smaller scores).

Then question pattern rules are applied. Some examples are shown in Figure 1 and there are 129 patterns in the working system (however, as there are many 'or's in the pattern, actual number would be very large). The main purpose of the pattern matching is to find NE type expected by the question. The types can be more than one, as shown by the first two rules. Each MATCH line is matched against each bunsetsu by Perl's pattern matching. NEXTBUNSETSU indicates that the matching proceeds to the next bunsetsu. There are two kinds of patterns. One is to find the NE type directly. The first four patterns are this pattern. For example, if the first pattern matches, NE type "organization" is proposed. The other type is that the pattern find 'center word' and the NE type is derived by the NE type specified by the word. The last two rules in the table is in this type. This is for the questions like "

„.

The last rule in Figure 1 is used and extract the word "　" as the center word. Then the system look up our center word dictionary, which specifies relationship between nouns and its NE types.

```
RULE start
RULEID DOKO-ORGANIZATION
MATCH    ^(     |    )
TYPE
SCORE    100.0
RULE end

RULE start
RULEID DOKO-LOCATION
MATCH    ^(     |    |    )
TYPE
SCORE    100.0
RULE end

RULE start
RULEID KEN-HA-DOKO-PROVINCE
MATCH    (  |  |         )  ?  $
NEXTBUNSETSU
MATCH    ^(     |   )(  |     |    |  |  )
TYPE
GENERALIZATION 0
PRIORITY 1
SCORE    10000.0
RULE end

RULE start
RULEID NANKASHO-N_LOCATION-0
MATCH       (  |  |  |  |  |  |  )*(  |  |  |
     |  )
TYPE
SCORE    1000.0
PRIORITY 1
RULE end

RULE start
RULEID XXX-HA-DOKO
MATCH [^      ]  $
NEXTBUNSETSU
MATCH ^(     |    )(  |     |    |  |  )
TYPEOF HEAD 1
PRIORITY 2
RULE end

RULE start
RULEID XXX-MEI-HA-NANI-PRE
MATCH (  |    )  $
NEXTBUNSETSU
MATCH    ^(  |     |    )(  |     |    |  |  )
TYPEOF PRE 1
PRIORITY 3
RULE end
```

Figure 1: Question pattern

Using the dictionary, the system finally figures out the expected NE type includes 'COMPANY'. The center word dictionary contains 16,431 entries and was compiled by hand based on Bunrui-Goi-Hyou and corpora. The rules have several attributes. PRIORITY to define the order of rule application, SCORE to indicate the likelihood of the NE type and GENERALIZATION to specify if the NE type can be generalized using the NE hierarchy.

Also, in the question examination component, keywords are identified. The keywords are used in both text retrieval and question extraction. Keywords include most kinds of nouns, adjectives, adverbs, verbs and unknown words. The keywords have scores based on POS type, IDF, and if it is center words or not. Keyword expansion is done using synonym dictionary, which contains 46,619 group of words, and the synonym words have relatively lower scores than the original words in the following processing. In the training phase, we figured out that having too many keywords is rather harmful, so the keywords are trimmed based on the score, number of keywords and overlappings to other keywords.

Several minor information is extracted, as well, which includes the context of interrogative word, the following word of interrogative in order to find suffix of number expressions (for example, "　　　" in "　　　　　　　　　　"), if the question is asking the definition of a word, if the question is asking alias. Such information is used in the various places of the following processing.

2.2　Text Retrieval

Text retrieval is basically done by something like Boolean search against paragraphs of articles (rather than the articles). In the search, the more kinds of keywords appears in the text, the more score the text gets. Only when the score is the same (i.e. the same number of kinds of keywords appears), the scores prepared in the question examination are used. We found, in the training phase, that there is an optimal number of text used in the following process. The text are deleted based on the number of text retrieved, absolute score difference to the top text, ratio difference to the top text.

2.3　Answer Extraction

The answers are extracted from the retrieved texts. The sentences in the texts were analyzed fully automatically by JUMAN and our NE taggers in advance. We used two NE taggers. One is Maximum Entropy based NE tagger using IREX's 8 NE definitions, trained by CRL-NE data. The

other is rule based system using 140 NE types, in which the NE dictionaries and rules are created by hand. The NE entities appeared in the previous paragraphs are also used in the answer extraction. The NE hierarchy is designed by our selves [Sekine et.al 2002] and available in the Web [NE Hierarchy].

The words in the retrieved texts which are tagged as nouns (except some special kinds of nouns), unknown words, NEs and sequence of nouns are taken as answer candidates. The system calculate scores for each candidate based on the distance from keywords, NE type, inclusion of center words, suffix, expression within brackets, if the question is asking alias and similarity of the context to the context in the question. We tried to use distance in terms of dependency, but we figured out the word distance performs better than the dependency distance using the training QA data.

2.4　Evaluation Result

The QAC task-1 evaluation result of our system reported in the messages from the task organizers on July 21 and August 7 is shown in Table 1. Here NQ# means Number of Questions which have correct answer in top #. We found a bit more than a quarter of correct answers in the top answer, and about a half in the top 5 answers. MRR was 0.39 which is 4th rank among the 15 participants.

Points	Answer	Output	Correct
75.6	305	1000	121
Recall	Precision	F-measure	MRR
39.67	12.10	18.54	0.39
RQ1	NQ1	RQ5	NQ5
0.297	58	0.523	102

Table 1: Our System Result

2.5　Error Analysis

We made error analysis which is shown in Table 2. First, the questions are categorized into four categories; 1) The top answer is correct (58), 2) The correct answer is found in top 2 to 5 (44), 3) The correct answer is not found in top 5 (93), 4) Answer does not exist (5). Then for the second and third categories, we analyzed the cause of errors in detail. Note that the total number of cause is more than the number of errors, as one error could include more than one causes.

In the second category, the answer extraction error is the largest. It basically means that ques-

Category	Freq.
Top Answer	**58**
Answer in 2-5	**44**
(2nd:23, 3rd:11, 4th:8, 5th:2)	
Crucial Juman error	2
Qpattern error	6
Keyword extraction error	5
NE tagger error	5
Answer extraction error	34
No Answer in top 5	**93**
Text retrieval error	27
Crucial Juman error	3
Center word error	6
Qpattern error	24
Keyword extraction error	18
NE tagger error	28
Answer extraction error	22
Answer does not exist	**5**

Table 2: Error Analysis of Our System

tion is analyzed correctly and the correct answer has correct NE tags, but the score of the correct answer is lower than that of incorrect ones. It includes the inevitable reason as long as we use simple word distance metric, where the incorrect word (which has the same NE to the correct answer) appears closer to the keywords than the correct word. A smarter method of including dependency distance may amend the problem. The frequencies of other causes are relatively small, but the qpattern error is a bit interesting. We made the question pattern based on more than 2000 QA training data, but the coverage seems not be enough. For example, questions " " and " " are the ones our rules could not cover. The first one was not considered just because of lack of coverage, and we were aware of the second type, but finding type based on the rare word like " " was very difficult (however, we have some coverage of such type of questions).

The causes of error in the category of "no correct answer was found in top 5" are rather mixed and are many time multiple. The number of 'text retrieval error' is 27. This means that texts which include the correct answers are retrieved for 168 questions out of 195 questions (86.15%). Most of the missing are caused by "keyword extraction error", which, we believe, happened at the keyword trimming using several criteria. Compared to that in the second category, NE tagger error are larger. It is because of heavy reliance on NE type in answer extraction. The NE errors include tagging errors to the following words, "

" (tagged as movie rather than car), " " (tagging span), " " (product rather than mineral), " " (no tag; suffix was not included), " " (vegetable rather than person), " " (separated), " " (no tag), " " (no tag), " " (no tag), " " (person rather than ship). The crucial Juman error are the followings. (the word border is indicated by "/"), " / / ", " / ", " / ", " (Noun)", " / ".

3 QAC Data Analysis

In this section, two surveys on the QAC data will be reported. One is type analysis of questions and the other is human performance on the QAC same questions.

3.1 Type Analysis

Table 3 shows the distribution of NE types of the QAC questions based on IREX's 8 categories. The classification based on 140 NE types is shown in Appendix (Table 5). Note that, in order to save space, the NE's which have same number and information are listed in the same row using ",," as delimiter, while keeping the order in the 140 NE table. From Table 3, we can see the distribution of NE types in QAC data compared with the distribution of NE types in general newspaper domain (from IREX NE task, formal evaluation, general domain). The percentage of person name is about the same, but LOCATION and ORGANIZATION is smaller in QAC data whereas ARTIFACT is much larger in QAC data. We believe it indicates that LOCATION and ORGANIZATION are not attractive enough to be used in QA, but ARTIFACT is. Also, another thought is that the same LOCATION and ORGANIZATION are mentioned many times compared with ARTIFACT. Except DATE, time and numeral expressions are about the same percentage in QAC data and IREX data. The reason that DATE is smaller might be the same as LOCATION and ORGANIZATION. Names, numbers and common nouns which are not covered by IREX categories occupies more than a quarter in QA data. This, we believe, is quite large. However there is almost no proper nouns which is not covered by our 140 NE categories. We still have to find how to deal with common noun type.

3.2 Human Performance

We asked two people to answer the QAC questions. The two people are both female, have uni-

IREX-NE	QAC	IREX
PERSON	42 (21%)	23.9%
LOCATION	29 + 1 (14.75%)	22.4%
ORGANIZATION	18 + 5 (10.25%)	27.4%
ARTIFACT	30 + 6 (16.5%)	3.2%
DATE	4 (2%)	17.2%
TIME	13 (6.5%)	3.6%
MONEY	3 (1.5%)	1.0%
PERCENT	3 (1.5%)	1.4%
Other (name)	22 (11%)	-
Other (number)	24 (12%)	-
Other (c. noun)	6 (3%)	-

Table 3: Answer Types Summary

versity level education and are frequent Web users. We first asked two people to answer the questions without looking at any information, then we asked them to answer them using Web in five minutes. The result is shown in Table 4. 'P1' and 'P2' in-

Who	Output	Correct
P1-w/o Web	88	61
P1-with Web	184	139
P1-with Web (edit)	184	168
P2-w/o Web	102	53
P2-with Web	188	140
P2-with Web (edit)	188	187
Best system (1)		98
Best system (5)		149

Table 4: Human Performance

dicate two people and, 'w/o Web' is the result without any information, 'with Web' is the result with five minutes Web access, 'with Web (edit)' is the result with five minutes Web access and the answers are edited when the human answer can be judged the same although the character string is not exactly the same. We had to make the third category, because we found that minor differences like inclusion of title (,), having brackets, different character type, minor variations of string. This is because the strict application of string matching against the answers prepared by the newspaper of limited time span is not fair for human. (However, the answer depending on time (for example, the baseball player who earns most between in 1998 and 1999 is different from the player in 2002) is not modified and is resulting mistake.)

The list of modifications for P2 is shown in Table 6 in Appendix. The first words in the pair is the one human made, and the second is the (clos-

est) correct answer QAC provides. We believe this is a good resource in the discussion how difficult to judge the correctness if we eliminate the restriction of "extracting" the answers from newspaper articles. We observed that the variation happens quite often, but the variations are not so irregular.

Comparing the human performance without Web, only top three systems outperform the best human if the answer is restricted to the best answer. If we allow system to have 5 answers, 10 systems performs better than human and 5 systems are worse than the best human. Using Web, the human performances are well better than the best system regardless to the number of answers 1 or 5. The questions both people can't find the answer at all are 1040, 1041, 1051, 1193 (There are more questions they could not find the correct answer.). Because the answer is guaranteed to be exist in the given newspaper articles (rather than no guarantees in Web in the human experiment), we believe we can do many things to get the human performance.

4 CRL QA Data

We created QA data using Mainichi 95 newspaper articles in the similar way to that of QAC data creation. It is publicly available at http://cs.nyu.edu/~sekine/PROJECT/CRL_QA/. There are three major differences to the QAC data.

1. it includes wider variety of questions including those without interrogative

2. answers are not checked by multiple systems, so the coverage may not sufficient

3. more information is included and the number of questions is larger (2000)

The information attached to each question-answer pairs are the followings.

- QUESTION: Question sentence

- Q_TYPE: Question type depending on interrogative and so on

- NE_TYPE: Named Entity type (140 NE; NE hierarchy version 4)

- CENTER_WORD: Center word, if exists

- LEVEL: Level of question, how ambiguous it is, if it need coreference etc

- ANSWER: Answer string

- DOCNO: ID of Document in which answer appears

Sample data is shown in Figure CRLQA in Appendix.

References

[IREX Committee 1999] IREX Committee 1999 *Proceedings of the IREX Workshop*

[Sekine and Isahara 2000] Satoshi Sekine, Hitoshi Isahara. 2000 : "IREX: IR and IE Evaluation Project in Japanese" *Proceedings of the LREC-2000 conference*

[Sekine et.al 2002] Satoshi Sekine, Kiyoshi Sudo, Chikashi Nobata. 2002 : "Extended Named Entity hierarchy" *Proceedings of the LREC-2002 conference*

[NE Hierarchy] NE Hierarchy Homepage `http://cs.nyu.edu/˜sekine/PROJECT/NEH`

[Juman] Juman Homepage (Juman Ver.3.61) `http://pine.kuee.kyoto-u.ac.jp/ nl-resource/juman.html`

Appendix

```
<QA>
<QAID>CRL-QA2002-00006-01</QAID>
<QUESTION>
                                </QUESTION>
<Q_TYPE> </Q_TYPE>
<NE_TYPE>    </NE_TYPE>
<CENTER_WORD>        </CENTER_WORD>
<LEVEL>A</LEVEL>
<A_SET>
<ANSWER>            </ANSWER>
<DOCNO>951111007</DOCNO>
</A_SET>
</QA>


<QA>
<QAID>CRL-QA2002-00344-01</QAID>
<QUESTION>


    </QUESTION>
<Q_TYPE>    </Q_TYPE>
<NE_TYPE>       </NE_TYPE>
<LEVEL>A</LEVEL>
<A_SET>
<ANSWER>       </ANSWER>
<DOCNO>951111190</DOCNO>
<DOCNO>951111079</DOCNO>
</A_SET>
</QA>


<QA>
<QAID>CRL-QA2002-01271-01</QAID>
<QUESTION>

</QUESTION>
<Q_TYPE>              </Q_TYPE>
<NE_TYPE>      </NE_TYPE>
<CENTER_WORD>    </CENTER_WORD>
<LEVEL>A</LEVEL>
<A_SET>
<ANSWER>            </ANSWER>
<DOCNO>951111077</DOCNO>
</A_SET>
</QA>


<QA>
<QAID>CRL-QA2002-01729-01</QAID>
<QUESTION>
                        </QUESTION>
<Q_TYPE>    </Q_TYPE>
<NE_TYPE>       </NE_TYPE>
<LEVEL>A</LEVEL>
<A_SET>
<ANSWER>     </ANSWER>
<DOCNO>951111253</DOCNO>
</A_SET>
</QA>
```

Figure 2: CRL QA data

	140-NE	IREX-NE
1		
42		PERSON
1		ORG
11		ORG
2		ORG
1		ORG
1		ORG
2		ORG
2		LOC
5		LOC
4	or	LOC
2		LOC
13		LOC
1	,	LOC
2		
1		
1		LOC,ARTI
1	or	ORG,ARTI
1		ORG, ARTI
2		ORG, ARTI
1	,	ORG, ARTI
11		ARTI
1		ARTI
2		ARTI
1		ARTI
2		ARTI
1	, ,	ARTI
1	, ,	ARTI
3		
1		ARTI
3	,	ARTI
2		ARTI
2		
2		
4		
1		
5		
4		DATE
13		TIME
1		
6		
3		MONEY
3		PERCENT
1		
3		
1	, ,	
3	,	
2		
1		
6		

Table 5: Answer Types in QAC

Table 6: Equivalent Human Answer