Proceedings of NTCIR-5 Workshop Meeting, December 6-9, 2005, Tokyo, Japan

Baseline Systems for NTCIR-5 CLQA1: An Experimentally Extended QBTE Approach

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

Abstract

This paper reports the performance of baseline systems for the NTCIR CLQA1 Task. As a task organizer for CLQA1 and hence a creator of both sample and formal run questions of JE/EJ subtasks, I have deemed it ideal to completely exclude the effect of human knowledge. Consequently, we have taken an approach to statistically construct baseline CLQA systems using only a QA data set. We employed the QBTE (Question-Biased Term Extraction) Model and a preliminary extended model of the QBTE Model, which statistically constructs QA systems only from question-answer pairs. The extended model uses word and POS dictionaries in addition to Q/A pairs. We constructed CLQA systems on the basis of a sample Q/A data set (300 Q/A pairs) that was provided by CLQA1 organizers. It took only a week to develop baseline CLQA systems, QATRO-JE/EJ/CE, but the results showed that 300 Q/A pairs is too small a number in a CLQA setting thus that only a small number of questions could be correctly answered by our system. An additional analysis after the formal run revealed that the cross-lingual setting makes it more difficult for the system to retrieve related documents and pin-point answers because of discrepancies between translated words and words actually appearing in questions and target articles.

Keywords: *NTCIR, Cross-Lingual QA, QBTE*

1 Introduction

One of the aims of the NTCIR CLQA1 (Pilot) Task is to evaluate the performance of Question Answering (QA) systems in the Cross-Lingual settings.

The framework of Cross-Lingual QA is for finding answers to a question in language X (*source language*) from documents in another language Y (*target language*). In NTCIR CLQA1, this is represented as the *XY subtask*. Since CLQA1 is a new pilot task in NTCIR, translating answers back to the source language is outside the scope of CLQA1.

Topics to be investigated are as follows.

- How to create Cross-Lingual QA (CLQA) systems between Asian languages
- The difference between a monolingual QA and cross-lingual QA
- The extent to which a CLQA system for the XY subtask degrades from the monolingual QA system in language Y
- The type of Machine Translation techniques that are effective for CLQA

As an organizer of CLQA1, I created the baseline CLQA systems QATRO [1] for JE/EJ/CE subtasks. To exclude an effect of my knowledge about CLQA test data, QATRO was built on a purely statistical method based on *QBTE (Question-Biased Term Extraction)* [11] and an experimentally extended QBTE Model, which only relies on Q/A data sets.

QBTE is a kind of Statistical Question Answering approaches [3, 4, 6, 7, 12, 13, 14, 10, 2, 5]. We employed the machine learning technique *Maximum Entropy Models (MEMs)* [1] to extract answers from combined features of question features and document features.

2 Overview

2.1 Subtasks Participated

We participated in the following two CLQA1 subtasks based on an extended QBTE Model.

JE subtask: We submitted three official and three unofficial runs.

[1] QATRO stands for Question Answering system with TRanslatiOn

EJ subtask: We submitted three official and three unofficial runs.

We participated in the following CE subtask using Chinese-to-English MT Engines and English QA system based on the QBTE Model.

CE subtask: We submitted three official and four unofficial runs.

2.2 Training Data

Document Sets: Japanese newspaper articles from the Yomiuri Shimbun Newspaper published in 2000 and 2001 and English article of The Daily Yomiuri published in 2000 and 2001.

Sample Question/Answer Set: We used CLQA1's sample QA data set. This dataset comprises of 300 JE/EJ-parallel questions with correct answers as well as question types and IDs of articles that contain the answers.

The document set is used not only in the training phase but also in the execution phrase.

2.3 Test Data

Formal Run Question The two hundred English, Japanese, and Chinese questions for the formal run, provided by the CLQA1 organizers.

2.4 Formal Runs

There are two kinds of runs evaluated in CLQA1:

Official Runs A set of the best answer and its DOCID for the 200 formal run questions.

Unofficial Runs A set of the five best answers and their DOCIDs for the 200 formal run questions.

3 Conventional Method: QBTE Model

This section briefly introduces a statistical QA framework, the QBTE Model, to construct a QA system from question-answer pairs based on the QBTE approach. When a user gives a question, the framework is to find answers to the question in the following two steps.

Document Retrieval retrieves the top N articles or paragraphs from a large-scale corpus.

QBTE creates input data by combining the question features and documents features, evaluates the input data, and outputs answers.

Table 1. EJ/JE POS dictionary

E-POS	J-POS
NN	名詞-一般
NNS	名詞-一般
NP	名詞-固有名詞
NPS	名詞-固有名詞
CD	名詞-数
JJ	形容詞
JJR	形容詞
JJS	形容詞
VV	動詞
VVN	動詞
VVD	動詞
VHZ	動詞
VHG	動詞
IN	助詞
CC	助詞-並立助詞
RB	副詞-一般
PP	名詞-代名詞-一般
TO	助詞
VBG	助動詞
VBN	助動詞
VBP	助動詞
WDT	関係詞
MD	助動詞

We used a simple idf method in document retrieval. Let w_i be words and w_1,w_2,$\ldots w_m$ be a document. Question Answering in the QBTE Model involves directly classifying words w_i in the document into answer words or non-answer words. That is, given input $x^{(i)}$ for w_i, its class label is selected from among $\{I, O, B\}$ as follows:

I: if the word is in the middle of the answer word sequence;

O: if the word is not in the answer word sequence;

B: if the word is the start word of the answer word sequence.

The class labeling system in our experiment is IOB2 [9], which is a variation of IOB [8].

There are three groups of features that can be used for features of input data:

- Question Feature Set (QF): Features extracted from a question;
- Document Feature Set (DF): Features extracted from a document;
- Combined Feature Set (CF): Combinations of question and document features.

In the QBTE Model, MEMs use only the Combined Feature Set.

Table 2. Submitted Runs

Run	Subtask	Output	Model	Training data	# of docs
QATRO-E-J-01	EJ	top 1	extended QBTE	300 sample Q/A	top 5
QATRO-E-J-02	EJ	top 1	extended QBTE	300 sample Q/A	top 20
QATRO-E-J-03	EJ	top 1	extended QBTE	300 sample Q/A	top 50
QATRO-E-J-u-01	EJ	top 5	extended QBTE	300 sample Q/A	top 5
QATRO-E-J-u-02	EJ	top 5	extended QBTE	300 sample Q/A	top 20
QATRO-E-J-u-03	EJ	top 5	extended QBTE	300 sample Q/A	top 50
QATRO-J-E-01	JE	top 1	extended QBTE	300 sample Q/A	top 5
QATRO-J-E-02	JE	top 1	extended QBTE	300 sample Q/A	top 20
QATRO-J-E-03	JE	top 1	extended QBTE	300 sample Q/A	top 50
QATRO-J-E-u-01	JE	top 5	extended QBTE	300 sample Q/A	top 5
QATRO-J-E-u-02	JE	top 5	extended QBTE	300 sample Q/A	top 20
QATRO-J-E-u-03	JE	top 5	extended QBTE	300 sample Q/A	top 50
QATRO-C-E-01	CE	top 1	SMT + QBTE	300 sample Q/A	top 20
QATRO-C-E-02	CE	top 1	SMT + QBTE	300 sample Q/A	top 50
QATRO-C-E-03	CE	top 1	EBMT + QBTE	300 sample Q/A	top 20
QATRO-C-E-u-01	CE	top 5	SMT + QBTE	300 sample Q/A	top 20
QATRO-C-E-u-02	CE	top 5	SMT + QBTE	300 sample Q/A	top 50
QATRO-C-E-u-03	CE	top 5	EBMT + QBTE	300 sample Q/A	top 20
QATRO-C-E-u-04	CE	top 5	EBMT + QBTE	300 sample Q/A	top 20

4 Preliminary Extended QBTE Model

For JE/EJ subtasks, we used the QBTE Model experimentally extended for Cross-Lingual QA. Given JE and EJ word dictionaries and a POS dictionaries (Table 1), QATRO-JE/EJ (Fig. 1) translate words and POSs to match question and document words based on the QBTE Model. Thanks to the Statistical QA approach, it took only one week to construct QATRO-JE/EJ/CE.

The numbers of word entries in the JE/EJ word dictionaries are as follows:

JE word dictionary: 24,805 entries,

EJ word dictionary: 17,571 entries.

For JE/EJ subtasks, we experimentally extended the model in the following two points.

- Document retrieval retrieves articles containing question words translated using a JE or EJ word dictionary.
- The Combined Feature Set includes combinations of words and POSs of a question and *translated* words and POSs of documents.

4.1 Extended CF

For each word w_i, the following features are created.

cw–k,. . .,cw+0,. . .,cw+k: matching results (true/false) between each of dw–k,...,dw+k features and any *translated* qw feature, e.g., *cw–1:true* if *dw–1:Christmas* and *qw: クリスマス (Christmas)*

cm1–k,. . .,cm1+0,. . .,cm1+k: matching results (true/false) between each of dm1–k,...,dm1+k features and any *translated* POS1 in qm1 features

cm2–k,. . .,cm2+0,. . .,cm2+k: matching results (true/false) between each of dm2–k,...,dm2+k features and any *translated* POS2 in qm2 features

cm3–k,. . .,cm3+0,. . .,cm3+k: matching results (true/false) between each of dm3–k,...,dm3+k features and any *translated* POS3 in qm3 features,

cm4–k,. . .,cm4+0,. . .,cm4+k: matching results (true/false) between each of dm4–k,...,dm4+k features and any *translated* POS4 in qm4 feature

cq–k,. . .,cq+0,. . .,cq+k: combinations of each of dw–k,...,dw+k features and *translated* qw features, e.g., *cq–1:President&When* is a combination of *dw–1:President* and *qw:*何時

5 Experimental Results

We participated in the JE/EJ subtasks and CE subtask. The results were evaluated using the Accuracy, the Top5 score, and MRR.

Accuracy denotes the rate at which the first ranked answers are correct.

Table 3. Official Results of EJ/JE Subtasks

Run	IR	R	R+U	Acc (R) (%)	Acc (R+U) (%)
QATRO-E-J-01	5	0	1	0.0	0.5
QATRO-E-J-02	20	0	0	0.0	0.0
QATRO-E-J-03	50	0	0	0.0	0.0
QATRO-J-E-01	5	2	2	1.0	1.0
QATRO-J-E-02	20	2	2	1.0	1.0
QATRO-J-E-03	50	1	1	0.5	0.5

Table 4. Unofficial Results of JE/EJ Subtasks

		R					R+U					R (%)			R+U (%)		
Run	IR	1	2	3	4	5	1	2	3	4	5	Acc	MRR	Top5	Acc	MRR	Top5
QATRO-E-J-u-01	JA	0	0	0	0	0	1	1	0	0	0	0.0	0.0	0.0	0.5	0.8	0.01
QATRO-E-J-u-02	JA	0	0	0	0	0	0	2	0	0	0	0.0	0.0	0.0	0.0	0.5	0.01
QATRO-E-J-u-03	JA	0	0	0	0	0	0	1	0	0	0	0.0	0.0	0.0	0.0	0.3	0.5
QATRO-J-E-u-01	EN	1	0	0	0	1	2	0	1	0	1	0.5	0.6	1.0	1.0	1.3	2.0
QATRO-J-E-u-02	EN	1	0	1	1	1	2	0	2	1	1	0.5	0.9	2.0	1.0	1.6	3.0
QATRO-J-E-u-03	EN	0	0	1	0	0	0	2	1	1	0	0.0	0.2	0.5	0.0	0.8	2.0

Top5 Score indicates the rate at which at least one correct answer is included in the top five answers.

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

5.1 Formal Run Results

Table 2 shows system parameters, while Table 3-6 present formal evaluation results of CLQA1 formal runs.

5.2 Additional Results

To investigate the difference between monolingual and cross-lingual QA, we conducted additional experiments. We created an English monolingual QA System (QATRO-E) and Japanese monolingual QA system (QATRO-J) with QBTE Model 1 using 300 sample Q/A pairs.

To evaluate QATRO-E, we used English formal-run questions for the EJ subtask and the English corpus for JE subtask, *i.e.*, The Daily Yomiuri newspaper articles. For QATRO-J, we used Japanese questions of formal run questions of the JE subtask and the Japanese corpus for the EJ subtask, *i.e.*, Yomiuri Shimbun newspaper articles..

Table 7 shows the performance of QATRO-E and QATRO-J.

6 Discussion

First of all, it is obvious that the number of training data, i.e., 300, was not sufficient, which means that we need more Q/A training data to improve performance.

Further, the Cross-Lingual setting makes QA tasks more difficult. Since translation from the source language to the target language is a kind of distortion, translated question words tend not to directly match document words. Because of this distortion effect, both document retrieval and QBTEs for CLQA become less accurate than monolingual QA. Moreover, the scores of Cross-Lingual QA systems became about half of monolingual QA systems.

7 Conclusion

We created baseline systems for CLQA1 JE/EJ/CE subtasks over the course of a week. The baseline systems QATRO-JE/EJ for JE/EJ subtasks were statistically constructed using 300 sample question/answers while the baseline system QATRO-CE for the CE subtask was a combination of MT engines and a statistically constructed monolingual QA system using 300 training question/answers.

The results indicate that Cross-Lingual setting degrades the performances of QA systems. To overcome this difficulty, we need more sample data and a more sophisticated extension of QBTE Model 1.

Acknowledgment

This research was supported by a contract with the National Institute of Information and Communications Technology (NICT) of Japan entitled, “A study of speech dialogue translation technology based on a large corpus”.

Table 5. Official Results of CE Subtask

Run	IR	R	R+U	Acc (R) (%)	Acc(R+U) (%)
QATRO-C-E-01	5	5	5	2.5	2.5
QATRO-C-E-02	20	3	3	1.5	1.5
QATRO-C-E-03	5	2	2	1.0	1.0

Table 6. Unofficial Results of CE Subtask

		R					R+U					R (%)			R+U (%)		
Run	IR	1	2	3	4	5	1	2	3	4	5	Acc	MRR	Top5	Acc	MRR	Top5
QATRO-C-E-u-01	5	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	20	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	5	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	20	2	1	1	5	2	2	1	1	5	2	1.0	2.2	5.5	1.0	2.2	5.5

References

[1] Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra: A Maximum Entropy Approach to Natural Language Processing, *Computational Linguistics*, Vol. 22, No. 1, pp. 39–71 (1996).

[2] Abdessamad Echihabi and Daniel Marcu: A Noisy-Channel Approach to Question Answering, *Proc. of ACL-2003*, pp. 16-23 (2003).

[3] Abraham Ittycheriah, Martin Franz, Wei-Jing Zhu, and Adwait Ratnaparkhi: Question Answering Using Maximum-Entropy Components, *Proc. of NAACL-2001* (2001).

[4] Abraham Ittycheriah, Martin Franz, Wei-Jing Zhu, and Adwait Ratnaparkhi: IBM's Statistical Question Answering System – TREC-10, *Proc. of TREC-10* (2001).

[5] Lucian Vlad Lita and Jaime Carbonell: Instance-Based Question Answering: A Data-Driven Approach: *Proc. of EMNLP-2004*, pp. 396–403 (2004).

[6] Hwee T. Ng, Jennifer L. P. Kwan, and Yiyuan Xia: Question Answering Using a Large Text Database: A Machine Learning Approach: *Proc. of EMNLP-2001*, pp. 67–73 (2001).

[7] Marisu A. Pasca and Sanda M. Harabagiu: High Performance Question/Answering, *Proc. of SIGIR-2001*, pp. 366–374 (2001).

[8] Lance A. Ramshaw and Mitchell P. Marcus: Text Chunking using Transformation-Based Learning, *Proc. of WVLC-95*, pp. 82–94 (1995).

[9] Erik F. Tjong Kim Sang: Noun Phrase Recognition by System Combination, *Proc. of NAACL-2000*, pp. 55–55 (2000).

[10] Yutaka Sasaki, Hideki Isozaki, Jun Suzuki, Kouji Kokuryou, Tsutomu Hirao, Hideto Kazawa, and Eisaku Maeda, SAIQA-II: A Trainable Japanese QA System with SVM, *IPSJ Journal*, Vol. 45, NO. 2, pp. 635-646, 2004. (in Japanese)

[11] Yutaka Sasaki, Question Answering as Question-Biased Term Extraction: A New Approach toward Multilingual QA, *in Proc. of ACL-2005*, pp. 215-222, Ann Arbor (2005).

[12] Jun Suzuki, Yutaka Sasaki, and Eisaku Maeda: SVM Answer Selection for Open-Domain Question Answering, *Proc. of Coling-2002*, pp. 974–980 (2002).

[13] Jun Suzuki, Hirotoshi Taira, Yutaka Sasaki, and Eisaku Maeda: Directed Acyclic Graph Kernel, *Proc. of ACL 2003 Workshop on Multilingual Summarization and Question Answering - Machine Learning and Beyond*, pp. 61–68, Sapporo (2003).

[14] Ingrid Zukerman and Eric Horvitz: Using Machine Learning Techniques to Interpret WH-Questions, *Proc. of ACL-2001*, Toulouse, France, pp. 547–554 (2001).

Table 7. Additional Results in Monolingual QA

		R					R+U					R (%)			R+U (%)		
Run	IR	1	2	3	4	5	1	2	3	4	5	Acc	MRR	Top5	Acc	MRR	Top5
QATRO-E	20	11	6	10	5	4	16	9	11	7	5	5.5	9.7	18.0	8.0	13.5	24.0
QATRO-J	20	4	1	2	2	2	9	5	6	4	7	2.0	3.0	5.5	4.5	8.0	15.5

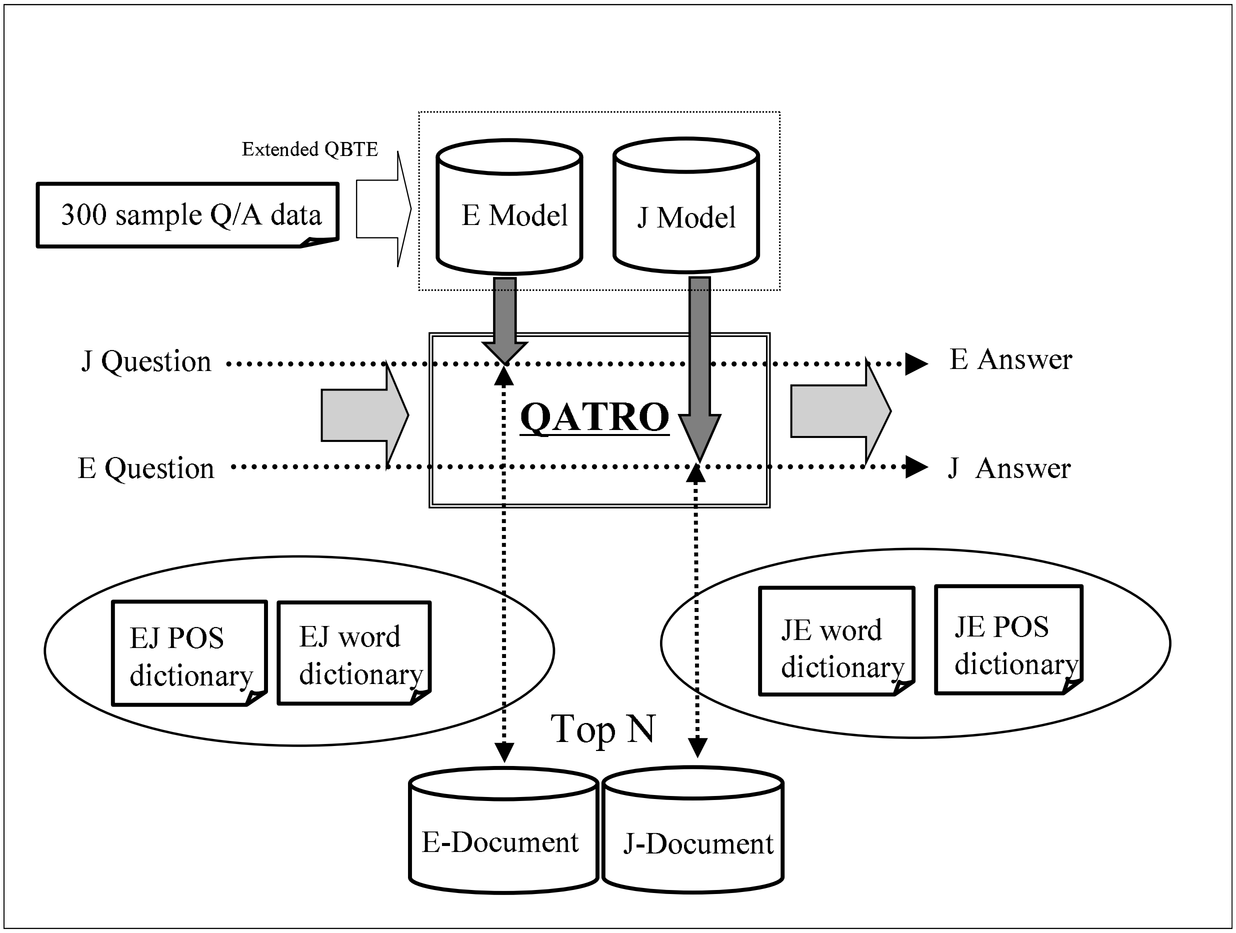

Figure 1. Block Diagram of QATRO-EJ/JE