System Description of NTOUA Group in CLQA1

Chuan-Jie Lin[1] Yu-Chun Tzeng[2] Hsin-Hsi Chen[2]
[1]Department of Computer Science
National Taiwan Ocean University
No 2, Pei-Ning Rd., Keelung 202, Taiwan
[2]Department of Computer Science and Information Engineering
National Taiwan University
No 1, Roosevelt Rd. Sec 4, Taipei 106, Taiwan
E-mail: cjlin@mail.ntou.edu.tw; hh_chen@csie.ntu.edu.tw

Abstract

CLQA1 is the first large scale evaluation on Chinese question answering. Our group participated in the C-E subtask. We augmented our monolingual Chinese QA system to handle cross-lingual QA. A bilingual dictionary and online web search engines were used to do the translation. Six runs were submitted at last, and the best run could provide correct answers of 8 of 200 questions at top 1 and 22 of 200 questions by top-5 answers.

Keywords: Chinese QA, cross-lingual QA, named entity translation

1. Introduction

Question answering has been a hot research topic in recent years. Since 1999, TREC starts to organize QA tracks [7], which provide environments of large scale evaluation in QA. Such resources are very valuable for QA researches.

NTCIR has already organized QA tracks since 2001 (QAC) [1] and starts a CLQA Track including JE/EJ/CC/EC/CE subtasks this year. We are very glad to know that a large scale evaluation on Chinese QA is now possible, and decided to participate in CE subtask.

We have started researches in question answering for a long time. Experiences from participating in TREC QA-Track for several years helped us to develop an English QA system [2]. After that, we applied our experiences to Chinese QA [3][5] and multimedia QA [3]. An online Chinese QA system has been established[1].

This year, we further extended our system to handle cross-lingual QA. Our first attempt was to receive Chinese questions then find answers in English documents. We participated in CE subtask in order to evaluate our system.

Figure 1 is the architecture of our cross-lingual QA system. After receiving a Chinese question, the question is first word-segmented, POS-tagged, and syntactically parsed. The information of its words, phrases, and syntactic structures will be used in following modules. Systems for word segmentation, POS-tagging, and parsing are developed in our lab.

The question type is decided by a question classifier described in Section 2, and the question core is extracted at the same time.

Words in the question except stop words are considered as its keywords and are used to form a query for an IR system to retrieve relevant documents from the CLQA collection. Our IR system uses Boolean model for QA purpose.

But before IR, the question should be translated into target language first. Translation module will be introduced in Section 3, as well as unknown word translation.

Named entities matching the question type are possible answer candidates. A NE identifier is used to identify occurrences of answer candidates in relevant documents. Every occurrence of each answer candidate is scored by scoring functions. After sorting by the scores, top-N candidates can be proposed as answers to this question. How to extract named entities and score them is illustrated in Section 4.

Section 5 gives the performance of our system in CE subtask with some discussions.

2. Question Classification

The question classifier we adopted for CE subtask was exactly the same one as we used in a monolingual Chinese QA system [5] since the questions were written in Chinese.

A question type is in fact referred to its answer type. If a question asks some person’s name, such as in the question “Who invented clips?” the question type (answer type) should be identified as PERSON. We defined 11 questions types, depending on their lengths and

[1] http://nlg.csie.ntu.edu.tw/

meanings. All question types are listed in Table 1.

Table 1. Question types

Boolean Answers
Yes-No, Selection
Short Answers
Person, Location, Time, Quantity, Object
Long Answers
Definition, Reason, Person Description, Method

A brief description of how we trained a question classifier [5] is given here: in order to learn question classification rules, we collected Chinese question sentences from a large corpus. There were 16,851 sentences terminated with question marks in the Academia Sinica Balanced Corpus. Lab members were asked to mark certain information in these question sentences, such as their types and the interrogative indicators which made the sentence "a question".

At first we adopted C4.5 to learn rules for classifying questions. But we found that the performance was very poor for some question types (e.g. TIME) Because these questions were selected from literal articles, not from a log of a real QA system, some types of questions did not occur quite often, and some questions were not even real questions, but to emphasize authors' opinions or emotions.

To improve the ability of question classifier, we expanded the rules learned by C4.5 to cover as many cases as possible. The modification procedures were conducted according to these principles:

(1) Syntactic Structures
Syntactic Structures were used to decide question cores. A question core is a certain phrase in the *what*-question which indicates its answer type. For example, in the question "哪個歐洲國家…" the word "國家" (country) hints that the answer should be a country name. In such a case, the word "國家" is the head of the noun phrase where the interrogative word appears.

(2) Synonyms
Two kinds of synonyms were added into the rules: synonyms of interrogative words and synonyms of question cores.
Synonyms of interrogative words were colleted from the Sinica Corpus, Cilin, and by experience. Synonyms of question cores were collected from Cilin. Different groups of words were adopted according to different question classes as examples shown in Table 2.

Table 2. Examples of question cores

QType	Heads in Question Core
Person	演員(actor), 總統(president), 選手(player) …
Location	國家(country), 城市(city) …
Time	年(year), 天(date) …
Quantity	高度(height), 長度(length)

(3) Part-of-Speech
Numbers are often seen in questions. In Chinese, there are also "quantifiers" following numbers. Different quantifiers will be used according to different nouns they modified. In order not to make a thorough list of words, POS information was used to detect the occurrences of numbers and quantifiers.

After modifying the question classification rules, the performance was improved. If only Person, Location, Time, and Quantity types were considered, the accuracy was improved from 58.3% to 88.7%. Examples of the new rules are given in Figure 2.

3. Question Translation

Translations of keywords in a question are used to (1) create a query to retrieve relevant documents; (2) detect their occurrences in relevant documents when scoring answer candidates (see Section 4).

person	{PERSON_cores} 是 [哪] 幾,一,POS=Neu 個,位,個人,POS=Nf, TAGWORD_Nflist
location	[什麼,甚麼,啥,何] {LOCATION_single},{LOCATION_cores}
time	[幾] 分,點,月,日,號
quantity	[幾] 天,年,時,期,TAGWORD_Nflist,POS=Nf,{QUANTITY_cores}, TAGWORD_OneNf,POS=Nf

Figure 2. Examples of Chinese question classification rules

3.1. Translation of known words

The translation was mainly done by dictionary lookup. We merged several English-Chinese dictionaries in order to obtain a better coverage [4]. For each word occurring in a question, if it could be found in the merged dictionary, we selected its first two English corresponding words to be its translations.

3.2. Translation of unknown words

Unknown words are mostly named entities. Translation of named entities is not an easy job. We proposed a method which consulted web search engines to find possible translations [6].

The basic algorithm was as follows. Top-*k* snippets returned by Google were analyzed. For each snippet, we collected those continuous capitalized words, and regarded them as candidates. Then we counted the total occurrences of each candidate in the *k* snippets, and sorted the candidates by their frequencies. The candidates of the larger occurrences were considered as the translation of the query term.

The above algorithm did not consider the distance between the query term and the corresponding candidate in a snippet. Intuitively, the larger the distance was, the less possible a candidate was. We modified the basic algorithm as follows. We dropped those candidates whose distances were larger than a predefined threshold. In this way, a snippet might not contribute any candidates. To collect enough candidates – say, *cnum*, we had to examine more than *k* snippets. Because there might not always be *cnum* candidates, we stopped collecting when maximum (*max*) snippets were examined. Finally, the candidates were sorted by scores computed as follows.

$$score(qt, c_i) = \frac{freq(c_i)}{2} - \frac{AvgDist(qt, c_i)}{3}$$

where $score(qt,c_i)$ denotes a score function of a query term qt and a candidate c_i,
$freq(c_i)$ denotes the frequency of c_i, and
$AvgDist(qt, c_i)$ denotes the average distance between qt and c_i.

In this way, we preferred those candidates c_i of higher occurrences with the query term qt and smaller average distances.

4. Answer Candidates and Scoring

The question types in the CE subtask this year were restricted to named entities. Therefore, we adopted an English NE identifier to find named entities in the relevant documents, and then chose those named entities which matched the question types as answer candidates. We chose LingPipe [2] to identify English named entities. But it could only detect person, location, and organization names. We created simple rules to extract English temporal and numerical expressions. Artifact names were not handled this year.

Every occurrence of each answer candidate to a question was scored according to the matching of some features:

(1) Question keywords
All words except stop words in a question were considered its keywords. Any matching of the keywords (including their inflections) contributed a score. Each keyword would contribute score only once, which was the largest score according to different scoring functions.
There were two kinds of final scores we would use to do ranking. One was called Boolean Score, which was defined as the number of matching question keywords. The other was Main Score, where named-entity keywords contributed higher scores *NEweight* than ordinary keywords, and some other discounting or additional contribution was considered as follows.

(2) Distances
We took distance of an answer candidate and a question keyword into account. The longer the distance of a keyword was, the less score it could contribute. The discounting was approximately proportional to the reciprocal of distance divided by 3.

(3) Synonyms of keywords
The matching of synonyms also contributed a score. We designed a weight *SNweight* to adjust its influence.

(4) Title information
If a question keyword appears in the title of a web page, it somehow contributes information through the whole document. For examples, the name of a laboratory is often the title of the lab's web pages. Therefore, if the matching of a keyword occurred in the title part, it contributed a constant score *TITLEweight* despite its distance to the answer candidate.

Some features adopted in Chinese monolingual QA system could not applied to this subtask:

(1) Phrasal matching

[2] http://www.alias-i.com/lingpipe/

To perform phrasal matching, we need a syntactic parser to parse every sentence considered relevant to the question. However, due to the time limit, we did not incorporate any syntactic parser for English texts in our system.

(2) Semantic matching of expressions

A number can be written in Arabic numbers or in English or Chinese words. A specific date can be expressed more than two different ways. Hence, we interpret the information of a temporal expression into a normal form (so as for a numerical expression). And then we can perform perfect matching of two expressions written in different forms, as well as the partial matching of two temporal expressions (such as "March, 1999" vs. "Mar. 26, 1999".)

However, we have not yet developed such a module for English texts. Temporal and numerical expressions were treated in the same way as other named entity types.

(3) Occurrences

The number of occurrences of an answer candidate can also be regarded as a supporting evidence of its being a correct answer. But if we would like to take occurrences into account, there must be a large collection which can provide abundant answering information. Web may be a good choice, but CLQA collection is not large enough. Therefore, we decided not to use the occurrence score.

5. Performance

We participated in CE subtask in CLQA1, NTCIR-5. We submitted six runs, three as official runs and three as unofficial runs. The common QA strategies among the six runs are:

(1) Top two English translation were selected for each question word
(2) Top two web translations were selected for unknown words, where *cnum* in the equation in Section 3.2 was set to 30, and *max* was set to be 500.
(3) Boolean score was the first key when ranking answer candidates, and Main score was the second ranking key.
(4) *TITLEweight* was set to be 0.6.

The individual setting of each run was:

[ntoua-C-E-01]
Answer candidates were extracted in top 15 relevant documents retrieved by using all English translations of question keywords as an IR query.

[ntoua-C-E-02]
Same as ntoua-C-E-01 except that redundant translations were dropped in the IR query.

[ntoua-C-E-03]
Answer candidates were extracted in top 30 relevant documents without redundant query words removal. While matching question keywords in the documents, matching of synonyms contributed a score where *SNweight* was set to be 1.

[ntoua-C-E-u-01]
Same as ntoua-C-E-02 but top 5 answers were reported.

[ntoua-C-E-u-02]
Answer candidates were extracted in top 30 relevant documents without redundant translations removal. Named entities contributed a score of 1.5 (*NEweight*).

[ntoua-C-E-u-03]
Answer candidates were extracted in top 50 relevant documents without redundant translations removal. Matching of synonyms contributed a score discounting by 0.6 (*SNweight*).

The accuracy and MRR scores of these six runs are listed in Table 3 and Table 4:

Table 3. Performances of official runs

Run		C-E-01	C-E-02	C-E-03
R	Top1	3	5	6
	Acc1	0.015	0.025	0.030
R+U	Top1	4	6	7
	Acc1	0.020	0.030	0.035

Table 4. Performances of unofficial runs

Run		C-E-u-01	C-E-u-02	C-E-u-03
R	Top1	5	6	7
	Top5	21	13	18
	Acc1	0.025	0.030	0.035
	MRR	0.053	0.042	0.056
R+U	Top1	6	7	8
	Top5	22	17	22
	Acc1	0.030	0.035	0.040
	Acc5	0.110	0.085	0.110
	MRR	0.058	0.053	0.065

where R means "correct answers", U means "correct answer not supported by the document", Acc1 is the accuracy of top-1 answers, Acc5 is the accuracy of top-5 answers, and MRR of top-5 answers.

Although the run ntoua-C-E-03 was evaluated as the second best among all the official runs, our overall performance was not good enough comparing to another cross-lingual QA subtask, EC subtask. Translating by dictionary look-up was probably not a good choice since QA needs more accurate information to do a better job.

Besides, the differences between scores of runs were too small to make a conclusion. Although that there was a tendency that more relevant documents were used to find answer candidates, more correct answers would be found.

The handling of named entities in the questions of CE subtask is harder than other subtasks, because the questions contain many Japanese names. These Japanese names are translated into Chinese which may or may not use the same Kanji (or Chinese characters), not to mention those names spelled in Katakana. If the Chinese translations are different from the original Japanese names, our method will fail to find translations of unknown words.

Moreover, the strategies to identify Japanese names, especially for personal names and location names, are different from the ones for Chinese. It seems we have to further develop a Japanese NE identifier in order to answer such questions.

6. Conclusion

We submitted six runs to the CE subtask in CLQA1, NTCIR-5. If only answers judged as correct (R) were considered, the best run correctly answered 8 of 200 questions at top 1, and 22 of 200 questions by top-5 answers.

It was our first attempt to do cross-lingual question answering. We applied our experiences in monolingual Chinese QA and CLIR to develop a cross-lingual QA system. But the performance was not good enough. More techniques should be integrated in the system in order to achieve more acceptable performance.

Reference

[1] Fukumoto, Jun'ichi, Kato, Tsuneaki and Masui, Fumito, "Question Answering Challenge for Five Ranked Answers and List Answers - Overview of NTCIR4 QAC2 Subtask 1 and 2," *Proceedings of NTCIR-4*.

[2] Lin, Chuan-Jie and Hsin-Hsi Chen (2001) “Description of NTU System at TREC 10,” *Proceedings of the Tenth Text REtrieval Conference* (*TREC 2001*), NIST Special Publication 500-250, pp. 406-411. Online available: http://trec.nist.gov/pubs/trec10/t10_proceedings.html.

[3] Lin, Chuan-Jie, Hsin-Hsi Chen, Che-Chia Liu, Ching-Ho Tsai and Hung-Chia Wung (2001) “Open Domain Question Answering on Heterogeneous Data,” *Proceedings of ACL Workshop on Human Language Technology and Knowledge Management*, pp. 79-85.

[4] Lin, Wen-Cheng and Chen, Hsin-Hsi (2003) "Merging Mechanisms in Multilingual Information Retrieval," *Advances in Cross-Language Information Retrieval: Proceedings of 3rd Workshop of the Cross-Language Evaluation Forum, Lecture Notes in Computer Science*, LNCS 2785, pp. 175-186.

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

[6] Tzeng, Yu-Chun (2005) *A Study on Multilingual Question Answering Systems*, Master thesis, National Taiwan University (in Chinese).

[7] Voorhees, Ellen (2002) “Overview of the TREC 2001 Question Answering Track,” *Proceedings of TREC-10*, pp. 42-51.

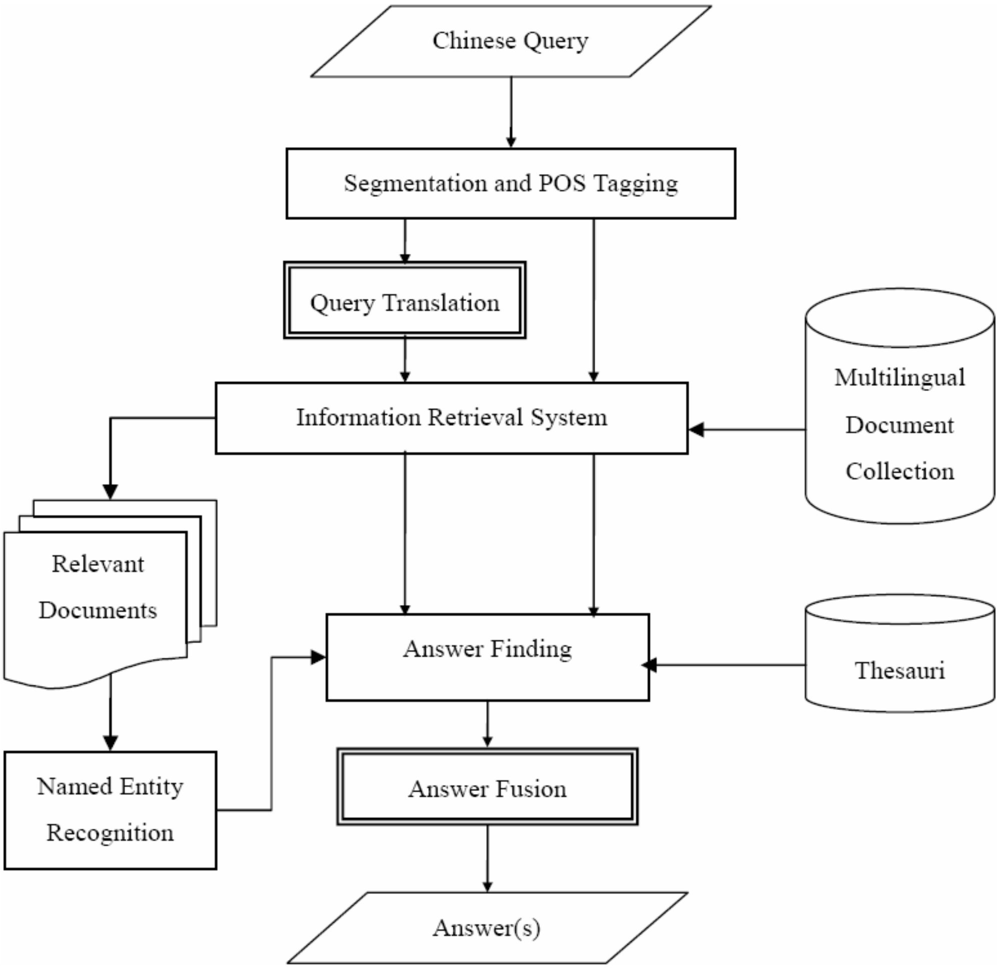

Figure 1. The architecture of the cross-lingual QA system of NTOUA