Two-Pass Named Entity Classification for Cross Language Question Answering

Yu-Chieh Wu
Department of Computer Science
National Central University
No. 300, Jhonda Rd., Jhongli City, Taoyuan County 32001, Taiwan
bcbb@db.csie.ncu.edu.tw

Kun-Chang Tsai
Department of Computer Science
National Central University
No. 300, Jhonda Rd., Jhongli City, Taoyuan County 32001, Taiwan
turninto@db.csie.ncu.edu.tw

Jie-Chi Yang
Graduate Institute of Network Learning Technology
National Central University
No. 300, Jhonda Rd., Jhongli City, Taoyuan County 32001, Taiwan
yang@cl.ncu.edu.tw

Abstract

In this paper, we present the mono-lingual and bilingual question answering experimental results at NTCIR6-CLQA. We combine most of the online resources and available resources to our QA systems without employing additional resources such as ontology, labeled data. Our method relies on three main important components, namely, passage retrieval, question classifier, and the named entity recognizer. Although our QA model is not state-of-the-art, the attractive of our method is that it was designed fully automatic without further adjusting the weights on different keywords. In the bilingual retrieval tasks, we translate the queries through a well-known machine translation tool. The evaluation results of our method were also given in the tail of this paper.

Keywords: Bilingual question answering, Chinese information retrieval, Chinese word segmentation, Named entity recognition, Question answering.

1 Introduction

The goal of question answering aims at acquiring the exact and short answer phrases in response to the user's questions. Usually, the question is quite different from the classical information retrieval, since the question is much more natural than the queries.

To construct a Q/A system, both IR and information extraction (IE) techniques are needed. The task of IE (Chieu and Ng, 2002; Bikel et al., 1999) is to recognize proper nouns and the named entity such as person name, location name, and organization names in text. It is the central theme in the Message Understanding Conferences (MUC). If we can classify the types of each question, e.g. people (who questions), location (where questions), date or time (when questions), we can then extract corresponding type of answers in texts. Generally speaking, most Q/A systems contain three components implicitly: question analyzer, document/passage retrievers, and answer finder (Suzuki et al., 2002). Given a question from users, the question analyzer will identify the question type. In the second step, important question terms will be used to retrieve related documents or passages. Finally the answer finder selects or ranks the possible answers among these candidate sentences or passages. Many Q/A systems (Lee et al., 2001; Suzuki et al., 2002; Tellex et al., 2003; Hovy et al., 2001; Wu and Yang, 2007) followed this basic idea and adopt complex and sophisticating methods to enhance the system performance.

In this paper, we focus on reporting the experimental results and system descriptions of our question answering model at NTCIR-6. The two important components, namely, IR and IE modules are adopted in our study. In this paper, we only focus on light-weight and efficient QA model construction, since most of the online resources such as named entity recognition, thesaurus (like WordNet) can not be ported to different domains

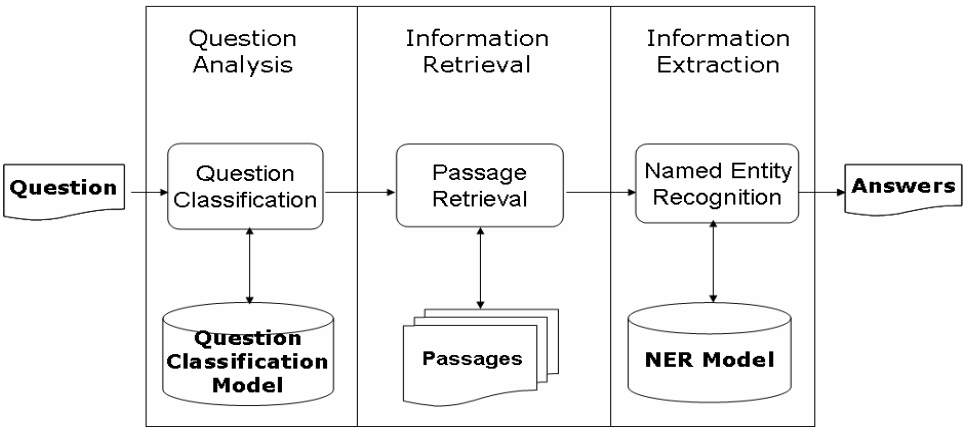

Figure 1: System architecture of the proposed question answering system

and languages, and also far away from the main centroid theme of the CLQA tasks.

The remainder of this paper is organized as follows: Section 2 describes the system overview of our method. In Section 3, the passage retrieval component is discussed. Section 4 and 5 outline the question classification and the named entity recognition modules, while in Section 6, we report the experimental results on the two tasks, monolingual and bilingual results. In Section 7, we draw the conclusion and future works.

2 System Overview

The system overview of our question answering system can be shown in Figure 1. As seeing the Figure 1, the most important three components are question classification, passage retrieval, and the final named entity recognition (NER). The question classification module sends the observed question class that restricts the NER tagger to extract the expected question type as answers. These components will be discussed in the following sections.

2.1 Machine Translation

In the bilingual question answering tasks, the query and collections may be in different languages. Thus, to make them consistently two translation strategies are frequently used, i.e., document translation, and query translation. The two strategies could apply the same translator for different purpose. One is to translate the whole document set, while the other simply converts the query sentences into another language. Usually, the query translation is much more tractable, and cheaper than the document translation since document set might take several GB size of storage spaces.

Original English Query	Singapore, Jan. 9 (CNA) The United States' strategy of direct engagement with mainland China improves bilateral relations and brings stability to East Asia, US political analyst Joseph Nye said. "The American strategy of engaging China as a responsible major power opens a future of benefits for the US, for China and for East Asia in general," he said. "It builds on current investments and strengths.It is alterable or reversible in the event that conditions change," he said. The dean of Harvard University's renowned John Kennedy School of Government described "engagement" as being more an attitude than a detailed policy. Professor Nye was giving a public lecture on "The Rise of China and the Future of International Security" to about 1,000 people, including government officials, academics and students, at the National University of Singapore.
Auto-translated into Chinese	新加坡， (CNA) 1 月 9 日直接訂婚美國的戰略與中國大陸改進雙邊關係并且給東亞帶來穩定，美國政治分析家約瑟夫 Nye 說。「參與中國美國戰略作爲負責任的主要力量打開未來好處爲美國，爲了中國和爲東亞」他一般來說，說。「它修造在當前投資和力量。它是可修改的或雙面布料在情況下情況改變，「他說。 政府哈佛大學的使有名望的約翰・肯尼迪學校的教務長比一項詳細的政策描述了「訂婚」作爲是更多態度。 Nye 教授在「中國的上升和未來給一個公開講演國際安全」關於 1,000 個人，包括政府官員，院和學生，在新加坡全國大學。

Figure 2: A sample of online web-page translation using systran web service

In this paper, we simply use several online translation tools[1] to translate the queries into the

[1] http://www.systranbox.com/
http://www.excite.co.jp/world/korean/
http://www.alphaworks.ibm.com/aw.nsf/html/mt

target languages. All of the queries were auto-converted into the specified target languages through these web sites. An example of text translation can be found in Figure 2. However, even these translation tools can successfully translate most of sentences, for some proper nouns, such as person names, project names, etc. it still need to be disambiguated. In this year, we do not employ the disambiguation methods to solve the proper noun translation problems since it requires some additional training corpus, and learners.

3 Passage Retrieval

Unlike traditional document retrieval, the target of NTCIR CLQA task is to extract the exact answer phrase (a series of words) that is able to answer the given question. To reduce the heterogeneity of the full document retrieval, the passage retrieval stands the intermediate roles between the document retrieval and the short answer extraction. Tellex et al. (2003) compared six well-known and famous passage retrievers for the TREC QA tasks. They showed that the BM-25 (Robertson et al., 2001; Savoy, 2005) retriever was slightly worse than the density-based models, included the SiteQ's, ISI, and IBM's methods which did combined many external resources. The focus of our QA approach is not the use of abundant materials to develop a state-of-the-art QA system since most of which can not be ported to different domains and languages, in particular to the NTCIR6-CLQA tasks. Thus, we created our own passage retrievers which can be established independently to the overall QA system constructions.

We directly replicate but slightly modify the BM-25 retrieval models to further speeding up the time efficiency in retrieval stage. However, different from most English-like languages, many Asian languages, such as Chinese, Japanese, and Korean do not have the space symbols to indicate the word boundaries. As reported in the other tasks of NTCIR, for example CLIR (Min et al., 2005; Chen et al., 2005; Savoy, 2005), the overlapping bigram level of words usually performs pretty well performance even though the high-performance word segmentation tool (Wu et al., 2006) is used.

In order to perform the passage retrieval, we firstly grouped three sentences with one overlapping as a paragraph. Then the conventional indexing and retrieval techniques are applied to those segmented "documents" (i.e., passages).

The well-known Okapi BM-25 weighting scheme is one of the top-performed ranking methods for document retrieval. More and more information retrieval studies further employed this method to Q/A systems and showed the effectiveness to retrieve documents and even the short passages (Tellex et al., 2003). However, to make the retrieval stage more efficiently, we re-write the BM-25 term weighting schema. Equation (1), (2), and (3) list the original BM-25 scoring functions.

$$\text{Document Score}(D) = \sum_{i=1}^{|Q|} W^{(1)} \frac{(k_1+1)tf(t_i,D)}{K+tf(t_i,D)} \frac{(k_3+1)tf(t_i,Q)}{k_3+tf(t_i,Q)} \quad (1)$$

$$W^{(1)} = \log\left(\frac{N-\mathrm{DF}(t_i)+0.5}{\mathrm{DF}(t_i)+0.5}\right) \quad (2)$$

$$K = (1-b)+b\times\frac{|\mathrm{D}|}{\mathrm{AVG}(|\mathrm{D}|)} \quad (3)$$

k_1, b, k_3 are constants, which empirically set as 1.2, 0.75, 500 respectively (Robertson et al., 2001; Savoy, 2005). $tf(t_i, D)$ represent the term frequency of term t_i in document D, and $tf(t_i, Q)$ represent the term frequency of term t_i in query Q. N denotes as the number of document in the collection. Equation (2) is merely a variant estimation of the "inverse document frequency" which can be observed after the all words were indexed. The third term in equation (1) should be computed in query stage which could be observed offline. However, the term $\frac{(k_1+1)tf(t_i,D)}{K+tf(t_i,D)}$ involves in measuring the average document length and the length of document D should be taken care. We now re-write equation (3) as follows.

$$\begin{aligned} K &= k_1((1-b)+b\times\frac{|\mathrm{D}|}{\mathrm{AVG}(|\mathrm{D}|)}) \\ &= k_1(1-b)+k_1\times|\mathrm{D}|\times b\times\frac{1}{\mathrm{AVG}(|\mathrm{D}|)} \\ &= c_1+c_2\times|\mathrm{D}| \end{aligned} \quad (4)$$

Where $c_1 = k_1(1-b)$ and $c_2 = k_1*b*1/\mathrm{AVG}(|D|)$. As mentioned above, k_1, b, k_3 were fixed constants that are unchanged during indexing, while c_2 is also a fixed constant after the statistics AVG(|D|) is found. By means of the above equations, we can efficiently compute the first term in equation (1)

via conventional indexing techniques. In this way, the first two terms of equation (1) can be stored in the indexed files. Once the query input to the retrieval system, the final ranked list is merely involved in computing the third term in equation (1) once and retrieving the indexed files. In our closed experiments, the retrieval time per query is less than 0.2 second since the computing of equation (3) at online stage is unnecessary.

4 Question Classification

The purpose of question classification is to detect the answer type of the input question. It shares the same target with conventional text categorization by treating a question sentence as a document and using various machine learning methods for classification. Amount of machine learning models had been presented to accomplish this task. Li and Roth (2002, 2005) provided 5000 manually annotated question set (UIUC question corpus[2]) and applied the SNoW algorithm to classify 500 TREC-10 testing questions into 50 for fine-grained and 6 for coarse-grained categories. In their works, they demonstrated that the additional human constructed word lists effectively improved the accuracy (from 79%~84.2%). Hacioglu and Ward (2003) reported that without using the word list, the support vector machine (SVM)-based question classifier can achieve competitive performance with SNoW (82%). In the same year, Zhang and Lee (2003) also showed that the use of parsing tree kernel for SVM, a better classification result for coarse-grained (six class) was obtained

As shown in previous studies combining with elaborated human-made word list, the improvement is quite effect.

Our question classifier is similar as traditional text classifier, which made use of bag of unigrams and bigrams as features. Unlike English question classification, there is no any famous benchmark corpus for this purpose. We thus developed our own labeled question set which mainly collected from the web logs, and blogs. The dataset consists of ~1000 Chinese questions and manually assigned to the predefined nine named classes as respective to the target of CLQA.

The classifier we adopted is support vector machines (SVM) (Joachims, 2001). SVM had been successfully applied to many classification problems such as pattern recognition and text categorization. However, the SVM is designed for binary categorization that should decompose multiclass into several of binary classes. We therefore extend it to multiclass SVM with conventional one-versus-all strategy.

5 Named Entity Recognition

Once the question classifier determines the question type, the goal of the following named entity recognition component (NER tagger) starts to identify the important answer phrase according to question type. Last year, our group had developed the high-performance Chinese word segmentation and named entity taggers by following the SIGHAN bake-off 3 shared task[3]. The tagger achieved almost 94% F-measure in the Chinese word segmentation results, while it 86% in F-measure in the Chinese named entity recognition task (Wu et al., 2006). This year, we further improve it with the SVM classifier and the same feature set as against to the conditional random fields that we adopted last year. More details can refer the two literatures (Wu et al., 2006; Lee and Wu, 2007).

Unfortunately, the SIGHAN corpus only covers three named classes as CLQA task, i.e., person, location, and organization names. Besides, the definition of the SIGHAN named entity task is quite different from CLQA since the data collection used in CLQA is mainly derived from the Taiwan United News, on the contrary, the annotated dataset of the SIGHAN named entity task came from Hong Kong news articles. The two different named definitions makes our NER tagger usually failed to capture the whole named chunk in the CLQA documents. One solution is to specifically develop a annotated named entity corpus to approximate the document collection of CLQA, since the domain adaptation is very another research issue and also out of scope of this paper.

In addition to the SIGHAN named entity corpus, we also combined another online resources to cover the other six un-supported named classes. The IEER entity extraction evaluation[4] was a simi-

[2] http://l2r.cs.uiuc.edu/~cogcomp/Data/QA/QC/

[3] http://sighan.cs.uchicago.edu/bakeoff2006/

[4] http://www.nist.gov/speech/tests/ie-er/er_99/er_99.htm

lar information extraction task as MUC-7 that was in Chinese. But the named and chunk definition between SIGHAN and IEER is also different with each other, we further developed the second NER tagger based on not only the original IEER corpus but also auto-tagged with the first NER tagger that was trained with the SIGHAN corpus. In other words, our NER tagger should perform two pass named recognition over the input text. Due to the time limitations, we have no sufficient time to evaluate our second NER tagger and the combinations of the two NER taggers. In the future, we plan to deeply test the actual performance of our NER tagger in the IEER benchmark corpus.

After the named entity recognition stage, the final answer is determined by choosing the closest pre-defined named chunk to the question word as the final answer. A better way for choosing the answer phrase could employ a question focus analysis technique to select the best named phrase instead of the ad-hoc selection. But the development of question focus analysis requires a rich labeled dataset, and also far away from the main purpose of this paper, we use the simple approach.

6 Experiments

The details about the query dataset and document collections were completely described in NTCIR-CLQA task definition [5]. The experiments were evaluated based on TREC-like procedure. Details can be found in the task definition of the NTCIR-CLQA session. In this year, we submit two runs of both monolingual QA and bilingual QA. The experimental results are explained as the follows.

6.1 Monolingual QA

The results for the monolingual question answering task in Chinese are listed in Table 2 and Table 3. As shown in Table 2, the performance of our question answering method only relies on the results of the three named entity categories, location, person, and organization, while the other six named types tend to be zero. We further observed that our could only handle the three categories due to the named entity recognition components always failed to extract the noun phrases of the other six types. For example, even the question classifier could identify the question type, the NER tagger still captures partial of the named phrase. This is the main shortcoming of our method on the question answering. The other six categories cover 47.34% of the questions. In other words, our method merely processed the 52.66% of the questions.

On the other hand, even the NER tagger could process those three name classes, our method still works well on the 52.66% of the questions.

Table 2: Monolingual C-C question answering results (Results of Right Answers)

NCUTW-C-C	Top1	Top5	MRR
ARTIFACT	0	0	0
DATE	0	0	0
LOCATION	0.2500	0.3750	0.2917
MONEY	0	0	0
NUMEX	0	0	0
ORGANIZATION	0.1875	0.1875	0.1875
PERCENT	0	0	0
PERSON	0.1277	0.2766	0.1922
TIME	0	0	0
All	0.0867	0.1467	0.1113

Table 3: Monolingual C-C question answering results (Results of Right+Unsupported Answers)

NCUTW-C-C	Top1	Top5	MRR
ARTIFACT	0	0	0
DATE	0	0	0
LOCATION	0.2500	0.3750	0.2917
MONEY	0	0	0
NUMEX	0	0	0
ORGANIZATION	0.2500	0.3125	0.2708
PERCENT	0	0	0
PERSON	0.1915	0.3617	0.2667
TIME	0	0	0
All	0.1133	0.1867	0.1436

6.2 Bilingual QA

Second, we continue the experiments to the bilingual question answering with only translating the English questions to retrieve Chinese collections. Table 4 lists the overall results on the E-C question answering task. Surprisingly, our method could additionally covers a little of the “DATE” type in this task. But it still mostly failed to answer the other six categories as the monolingual QA task. In

[5] http://clqa.jpn.org/

this experiment, our method still only works on the "PERSON", "LOCATION", and "ORGANIZATION" named classes.

Table 4: Bilingual E-C question answering results (Results of Right+Unsupported Answers)

NCUTW-E-C	Top1	Top5	MRR
ARTIFACT	0	0	0
DATE	0.0256	0.0256	0.0256
LOCATION	0.1250	0.1771	0.2500
MONEY	0	0	0
NUMEX	0	0	0
ORGANIZATION	0.0625	0.0781	0.1250
PERCENT	0	0	0
PERSON	0.0426	0.1284	0.2553
TIME	0	0	0
All	0.0400	0.0741	0.1267

7 Conclusion

This is our first time to participate the cross language question answering task. We mainly developed a Chinese named entity recognition system via integrating several online available resources and SIGHAN bake-off tasks. Our named entity recognizer was designed based on two-pass recognition for the two heterogeneous training datasets. We observe that our NER tagger can work well especially to some of the proper name classes, like person name, locations, and organization names. However, for the other categories, like date, time, etc. our method easily failed to extract the complete named phrase, instead, it only captures partial of the names. We expected that this could be easily recovered by relaxing the recognized named chunks.

In the bilingual question answering tasks, we here used the online machine translation tool to automatically convert one language into another. We also found that many disambiguation or word miss were mainly caused by the proper nouns, like person names, location names, etc. In the future, we plan to resolve the proper noun disambiguation problems by adopting a phrase chunker and some parallel linguistic corpora. To improve our QA systems, we will focus on developing much more homogeneous labeled dataset for our NER taggers.

References

D. Bikel, R. Schwartz, and R. Weischedel, "An algorithm that learns what's in a name," Machine Learning, pp. 211-231, 1999.

H.L. Chieu, and H.T. Ng, "Name entity recognition: a maximum entropy approach using global information," International Conference on Computational Linguistics (COLING), pp. 190-196, 2002.

J. Chen, R. Li, and F. Li, "Chinese information retrieval using Lemur," Proceedings of the 5th NTCIR workshop meeting on evaluation of information access technologies, 2005.

K. Hacioglu, and W. Ward, "Question classification with support vector machines and error correcting codes," In Proceedings of Human Language Technology Conference (HLT-NAACL), pp. 28-30, 2003.

E. Hovy, U. Hermjakob, and C.Y. Lin, "The use of external knowledge in factoid QA," In Proceedings of the 10th Text Retrieval Conference, pp. 644-652, 2001.

T. Joachims, "A statistical learning model of text classification with support vector machines," In Proceedings of the 24th ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 128-136, 2001.

G.G. Lee, J. Seo, S. Lee, H. Jung, B.H. Cho, C. Lee, B.K. Kwak, J. Cha. D. Kim, J. An, H. Kim, and K. Kim, "SiteQ: engineering high performance QA system using lexico-semantic pattern matching and shallow NLP," Proceedings of the 10th text retrieval conference, pp. 437-446, 2001.

Y.S. Lee, and Y.C. Wu, "A robust multilingual portable phrase chunking system," Expert Systems with Applications, vol. 33, no. 3, pp. 1-26, 2007.

X. Li, and D. Roth, "Learning question classifiers," In Proceedings of the 19th International Conference on Computational Linguistics, pp. 556-562, 2002.

X. Li, and D. Roth, "Learning question classifiers: the role of semantic information," *Journal of Natural Language Engineering, 2005.*

J. Min, L. Sun, and J. Zhang, "ISCAS in English-Chinese CLIR at NTCIR-5," Proceedings of the 5th NTCIR workshop meeting on evaluation of information access technologies, 2005.

E. Robertson, S. Walker, and M. Beaulieu, "Experimentation as a way of life: Okapi at

TREC," Information processing & management, vol. 36, no. 1, pp. 95-108, 2000.

Y. Sasaki, "Question answering as question-biased term extraction: a new approach toward multilingual QA," In Proceedings of the 43rd Annual Meeting of the ACL, pp. 215-222, 2005.

J. Savoy, "Comparative study on monolingual and multilingual search models for use with Asian languages," ACM trans. on Asian language information processing, vol. 4, no. 2, pp. 163-189, 2005.

J. Suzuki, Y. Sasaki, and E. Maeda, "SVM answer selection for open-domain question answering," In 19th International Conference on Computational Linguistics, pp. 974-980, 2002.

S. Tellex, B. Katz, J.J. Lin, A. Fernandes, and G. Marton, "Quantitative evaluation of passage retrieval algorithms for question answering," In Proceedings of the 26th ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 41-47, 2003.

Y.C. Wu, J.C. Yang, and Q.X. Lin, "Description of the NCU Chinese word segmentation and named entity recognition system for SIGHAN Bakeoff 2006," Proceedings of the 5th SIGHAN Workshop on Chinese Language Processing, pp. 209-212, 2006.

Y.C. Wu, and J.C. Yang, "Toward Multimedia: A String Pattern-based Passage Ranking Model for Video Question Answering", In Proceedings HLT-NAACL 2007, in press.

D. Zhang, and W.S. Lee, „Question classification using support vector machines," In Proceedings of the 26th ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 26-32, 2003.