Chinese-Chinese and English-Chinese Question Answering with ASQA at NTCIR-6 CLQA

Cheng-Wei Lee[12], Min-Yuh Day[1], Cheng-Lung Sung[1], Yi-Hsun Lee[1], Tian-Jian Jiang[12], Chia-Wei Wu[1], Cheng-Wei Shih[1], Yu-Ren Chen[1], Wen-Lian Hsu[1§]

[1]Institute of Information Science, Academia Sinica, Taiwan, R.O.C
[2]Department of Computer Science, National Tsing-Hua University, Taiwan, R.O.C
§corresponding author

{aska,myday,clsung,rog,tmjiang,cwwu,dapi,yrchen,hsu}@iis.sinica.edu.tw

Abstract

For NTCIR-6 CLQA, we improved our question answering system ASQA (Academia Sinica Question Answering System), which participated in NTCIR-5 CLQA, so that it could deal with the Chinese-Chinese (C-C) subtask and the English-Chinese (E-C) subtask. There are three innovations in the improved system: (a) to handle the E-C subtask, we have built an English question classifier that adopts Question Informer as a key classification feature; (b) with automatically generated Answer Templates, we can accurately pinpoint the correct answers for some questions. When Answer Templates are applied, the RU-accuracy is 0.911 for the applied questions; and (c) the Answer Ranking module has been improved by incorporating a new feature called, SCO-QAT (Sum of Co-occurrence of Question and Answer Terms). In NTCIR-6 CLQA, ASQA achieved 0.553 RU-accuracy in the C-C subtask and 0.34 RU-accuracy in the E-C subtask.

Keywords: *Question answering (QA), question classification, Question Informer, SCO-QAT, Answer Template*

1. Introduction

Because of the high level of information overload on the Internet, research into question answering, which focuses on how to respond to users' queries with exact answers, is becoming increasingly important. In recent years, many international question answering contests have been held at conferences and workshops, such as TREC [4], CLEF [1], and NTCIR [3]. Our proposed system, the Academia Sinica Question Answering System (ASQA), participated in the NTCIR-5 CLQA C-C subtask, and achieved 44.5 RU-accuracy. In NTCIR-6 CLQA, we used an enhanced version that incorporates three innovations: (a) to handle the E-C subtask, we built an English question classifier that adopts Question Informer as an important classification feature; (b) with automatically generated Answer Templates, we are able to accurately pinpoint the correct answers for some questions such that the RU-accuracy is 0.911 when the templates are applied; and (c) the Answer Ranking module has been improved by incorporating a new feature called SCO-QAT (Sum of Co-occurrence of Question and Answer Terms). In NTCIR-6 CLQA, we achieved 0.553 RU-accuracy in the CC subtask and 0.34 RU-accuracy in the EC subtask.

Hereafter, we refer to the original ASQA system used in NTCIR-5 as ASQA1, and the second version used in NTCIR-6 as ASQA2.

The remainder of this paper is organized as follows. Section 2 describes the system architecture. In Section 3, we introduce the three innovations incorporated in ASQA2. A detailed performance analysis of the innovations is reported in Section 4. Finally, we present a discussion and our conclusions in Section 5.

2. System Description

The architecture of ASQA2 (Figure 1) is the same as that of ASQA1 [7], except that the Answer Extraction module is divided into two modules: Answer Extraction and Answer Filtering. In addition, while ASQA1 can only answer Chinese questions, ASQA2 can deal with both mono-language and cross-language QA. We describe the innovations added to ASQA2 in the following subsections.

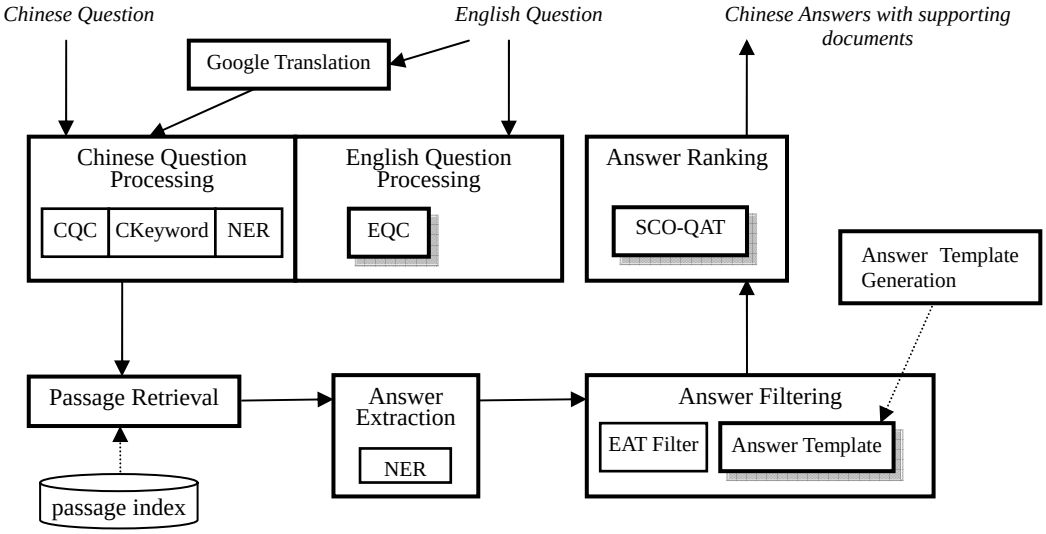

Figure 1. System architecture of ASQA2 for Chinese-Chinese and English-Chinese Factoid QA

Details of our three innovations are given in Section 3.

2.1. Chinese-Chinese QA

To deal with Chinese-Chinese QA (C-C QA), a question is first analyzed by the Chinese Question Processing module to obtain question types, keywords, QFocus, and NEs. Based on the keywords, the Passage Retrieval module retrieves related passages $\{P_1, P_2, P_3, \ldots, P_m\}$, where m $\leqq$ 100. In the Answer Extraction phase, candidate answers $\{A_{i1}, A_{i2}, \ldots, A_{in}\}$ are extracted from each passage, P_i, by a fine-grained NER engine. To eliminate irrelevant candidate answers, we employ the EAT (Expected Answer Type) Filter, which operates according to a mapping table that determines the compatibility of question types and answer types.

In addition to the above parts of ASQA2, which are identical to those of ASQA1 [7], we propose two new techniques, *Answer Template* and *SCO-QAT*, for pinpointing answers more accurately. Answer Templates are syntax patterns of question terms and answer terms. When an Answer Template matches some terms in a passage, it indicates there is a relation(s) between these terms. The templates comprise the core of the Answer Template Filter sub-module, responsible for filtering out candidate answers missed by the EAT Filter. Finally, pairs of candidate answers and their supporting passages, i.e., $PA\_Pairs = \{(P_1, A_{11}), (P_1, A_{12}), (P_1, A_{13}), \ldots, (P_2, A_{21}), (P_2, A_{22}), \ldots, (P_3, A_{31}), (P_3, A_{32}), \ldots..\}$, are ranked by the Answer Ranking module. The module has been enhanced to include three new features, namely, *answer frequency*, *passage score*, and *SCO-QAT*. All the ranking features are combined as a weighted-sum; the weights were originally trained on NTCIR5 CLQA data with a Genetic Algorithm.

2.2. English-Chinese QA

To deal with English questions, we incorporated Google Translate [2] into ASQA2 to translate English questions into Chinese. Although it is possible to feed the translated question directly to the C-C QA flow described in Section 2.1, the performance result of our experiments was not ideal. To resolve the problem, we have introduced an English Question Classification sub-module for more accurate identification of question types, which in turn improves answer filtering. Except for question type classification, other portions of the E-C flow are identical to those of the C-C flow.

3. Innovations

In this section, we describe the innovations added to ASQA2, namely, the English Question Classification sub-module, the Answer Template Filter sub-module, and the SCO-QAT answer ranking feature.

3.1. Classification of English Questions

Question Informer plays a key role in enhancing question classification for factual question answering [6]. In a previous work, to enhance Question Informer prediction, we proposed a hybrid approach that integrates a Genetic Algorithm (GA) with Conditional Random Fields (CRF) to optimize feature subset selection in a CRF-based model [5]. Krishnan et al. introduced the notion of Question

Informer for question classification and showed that human-annotated Question Informers lead to substantial improvements in the accuracy of question classification. They suggested choosing a minimal, appropriate contiguous span of a question token, or tokens, as the Question Informer of a question, which is adequate for question classification. For example, in the question: "What is the biggest city in the United State?" the Question Informer is "city". Thus "city" is the most important clue for question classification.

We use a machine learning approach, which is based on SVM classification, for English question classification. The training dataset for English question classification used in NTCIR-6 CLQA was based on our related work [Day et al., 2006]. We use Li and Roth's UIUC QC dataset [Li and Roth, 2002] and the corresponding Question Informer dataset from Krishnan et al. [6] to train the classification model. There are 5,500 training questions, 500 test questions, and corresponding question informers. Li and Roth used supervised learning for question classification of the UIUC QC data set; this is now the standard dataset for question classification [Day et al., 2006]. It has 6 coarse-grained and 50 fine-grained answer types in a two-level taxonomy, together with 5,500 training questions and 500 test questions.

We derived 4,204 valid questions tagged with their question types for CLQA factoid question answering. The questions were obtained from 6,000 UIUC questions with Question Informers by mapping the UIUC types to the ASQA question types. The question type taxonomy for English question classification includes 6 coarse-grained classes and 62 fine grained classes – the same as Chinese question classification in ASQA1 [Day et al., 2005]. We used an SVM model trained from 5,288 questions (ModelQ5288E: 4,204 questions from UIUC + 500 questions from NTCIR-5 CLQA development set + 200 questions from NTCIR-5 CLQA test set + 384 questions from TREC2002 500 questions) for English question classification of NTCIR-6 CLQA English questions. Note that we used different features (including Question Informer) to construct the SVM model based on a total of 5,288 English questions and their labeled question types.

3.2. Answer Filtering with Answer Templates

Answer Templates in ASQA2 are syntax patterns for identifying relations. The identified relations are then used to measure the correctness of an answer. This is similar to the concept of surface patterns used in several QA research projects [8, 9]. However, unlike surface patterns, Answer Templates do not target a particular question type. They are automatically generated and selected from training data for any kind of question type, and have the ability to capture important relations between a question's terms and the answer.

The Answer Template Filter sub-module utilizes the relations captured by Answer Templates to find relevant answers and filter out irrelevant ones. Compared to the EAT Filter, the Answer Template Filter is better able to identify correct answers. In fact, when an Answer Template is applied, only the best answer and its supporting passages are retained. If a template cannot be applied, it means there is no confident relation for the Answer Template Filter to identify correct answers; thus, all answers will be retained. In the following sub-sections, we describe how Answer Templates are created automatically, and how we use them to filter answers.

3.2.1. Answer Template Generation

We used local alignments of sentences to generate templates. Because sentence alignment is time consuming, instead of using the whole corpus, we only used the passages of training questions. They consisted of 400 NTCIR-5 CLQA questions, and 465 questions that we created. For each training question, only the top 200 passages returned by the Passage Retrieval module were collected and tagged with NE and POS tags. To align two sentences, we need a similarity function to determine the degree of similarity between two words. The similarity function is defined as

$$d(a,b) = \max \begin{cases} 1, & a = b \\ 1, & NE(a) = NE(b) \\ 1, & POS(a) = POS(b) \\ 1 - penalty, & POS(a) \approx POS(b) \\ 0, & a \neq b, \end{cases}$$

where *NE(a)* is the Named Entity (NE) tag of *a*, and *POS(a)* is the POS tag of *a*. If the POS tags of *a* and *b* are not the same, but they have a common prefix, the degree of similarity is subtracted with a penalty.

Our template generation (TG) algorithm extracts general patterns using the proposed alignment algorithm. We begin by pairing all the passages according to their similarity. Closely matched pairs are then aligned and a template that fits the passages of an aligned pair is created. A template is composed of ordered slots, which are chosen according to the corresponding parts of the aligned sentence pair with the following priority: word > NE tag > POS tag. If the sentences for a given slot have nothing in common, the TG algorithm creates a gap ("—") in that position. The generated templates are then processed in the template selection stage, described in the following section, to select Answer Templates.

Table 1. Answer Template application example. There two passages from udn_xxx_19991103_0700, two Answer Templates, two temporary relations, and a final relation.

Question:	女演員/OCC 蜜拉索維諾/PER 獲得/VJ 奧斯卡/Nb/ORG 最佳/A 女配角/OCC 獎/Na 是/SHI 因/Cbb 哪/Nep 部/Nf 電影/Na
$Passage_1$	..... 而/Cbb 奪得/VC 一九九五/Neu 奧斯卡/Nb 最佳/A 女配角/OCC 的/DE 殊榮/Na …
$Template_1$:	*VC Neu Nb A OCC - Na*
$Relation_1$:	{奪得/VC, 奧斯卡/Nb, 女配角/OCC}
$Passage_2$	… 蜜拉索維諾/PER 在/O/P/O 「/O/PAR 非強力春藥/ART 」/PAR 中/Ncd ..... 獲/VJ 奧斯卡/Nb 獎/Na …
$Template_2$:	*PER P PAR ART PAR – DE Na X VJ Nb*
$Relation_2$:	{蜜拉索維諾/PER, 非強力春藥/ART, 獲/VJ, 奧斯卡/Nb}
$Relation_3$:	{ 奪得/VC, 奧斯卡/Nb, 女配角/OCC, 蜜拉索維諾/PER, 非強力春藥/ART, 獲/VJ }

3.2.2. Answer Template Selection

To select useful answer templates, we start by applying generated templates to the training set. We extract important terms (i.e., terms with NEs and terms with the POS tag 'Nb' or 'V') from each training question and use these terms to fill the slots of the corresponding NE/POS tags. Next, the slot-filled templates are applied to all the passages selected for each question. If a slot-filled template matches a passage and the matched segment contains the correct answer to the question, the template is selected as an Answer Template.

3.2.3. Answer Template Matching

To filter candidate answers, we identify relations by matching the passages retrieved for a question with Answer Templates, and then calculate a score for each candidate answer based on the relations. If a template matches a passage, we extract a relation, which consists of the key matched terms (i.e., we discard terms that do not belong to an Nb, an NE, or a verb). If two relations contain overlapping terms (i.e., the same term is matched by at least two templates,) we check the *idf* values of the terms. If at least one of the *idf* values is higher then a given threshold, the two relations are merged. For example, the application shown in Table 1 contains a question, two retrieved passages from a document, and two templates that match the two passages. The first template, $Template_1$, extracts $Relation_1$ {奪得/VC, 奧斯卡/Nb, 女配角/OCC} from $Passage_1$, while $Template_2$ extracts the terms「蜜拉索維諾/PER」, 「非強力春藥/ART」,「獲/VJ」,「奧斯卡/Nb」and forms $Relation_2$. Since 「奧斯卡」 already exists in $Relation_1$, we examine the *idf* value of 「奧斯卡」 and merge it with $Relation_1$ to form $Relation_3$. After all the relations have been constructed for the given question, we use the question's key terms (女演員, 蜜拉索維諾, 獲得, 奧斯卡, 女配角 in this example) to filter out impropriate answers. If relations do not have any question key terms, we discard the candidate answers they contain.

Next, we calculate the score of each candidate answer according to the scores of the relations. A relation score is defined as the ratio of the question's key terms to the matched terms found in the relation. For example, in Table 3, the number of key terms in the question is 5, and the number of matched terms in $Relation_3$ is 3; thus, the score of $Relation_3$ is 3/5. After processing all the passages, we rank the candidate answers by the sum of the scores for the relations in which they appear, and retain the top ranked answer.

3.3. SCO-QAT Answer Ranking Feature

The basic assumption of SCO-QAT is that, in good quality passages, the more often an answer co-occurs with question terms, the higher the probability that it is correct. Next, we describe the SCO-QAT function. Assume the given answer is *A* and the given question is *Q*, which consists of a set of question terms *QT* $\{qt_1, qt_2, qt_3, \ldots\ldots, qt_n\}$. Based on *QT*, we define QC as a set of question term combinations, or more precisely $\{qc_i \mid qc_i$ is a non-empty subset of *QT* $\}$. We also define a *freq(X)* function of a set *X* to indicate the number of retrieved passages in which all elements of *X* co-occur. The confidence of a relation is calculated by:

$$Conf(qc_i, A) = \begin{cases} \dfrac{freq(qc_i \cup A)}{freq(qc_i)}, & \text{if } freq(qc_i) \neq 0 \\ 0, & \text{if } freq(qc_i) = 0 \end{cases} \quad (1)$$

Then, the SCO-QAT formula is defined as:

$$SCO\text{-}QAT(A) = \sum_{i=1}^{|QC|} Conf(qc_i, A). \quad (2)$$

For example, given a question Q consisting of three question terms {qt1, qt2, qt3 } and a corresponding answer set {c1, c2}, the retrieved passages are:

P1: qt1 qt2 c2
P2: qt1 qt2 qt3 c1
P3: qt1 qt2 c1
P4: qt1 c2
P5: qt2 c2
P6: qt1 qt3 c1

We use Equation (2) to calculate the candidate answer's SCO-QAT score.

$$\begin{aligned} SCO\text{-}QAT(c1) &= \frac{freq(qt1 \cup c1)}{freq(qt1)} + \frac{freq(qt2 \cup c1)}{freq(qt2)} \\ &+ \frac{freq(qt1 \cup qt2 \cup c1)}{freq(qt1 \cup qt2)} + \frac{freq(qt1 \cup qt3 \cup c1)}{freq(qt1 \cup qt3)} \\ &+ \frac{freq(qt1 \cup qt2 \cup qt3 \cup c1)}{freq(qt1 \cup qt2 \cup qt3)} \\ &= \frac{3}{5} + \frac{2}{4} + \frac{2}{3} + \frac{2}{2} + \frac{1}{1} = 3.77 \end{aligned}$$

$$SCO\text{-}QAT(c2) = \frac{2}{5} + \frac{2}{4} + \frac{1}{3} + \frac{0}{2} + \frac{0}{1} = 1.23$$

Since the SCO-QAT score of c1 is higher than that of c2, c1 is considered a better candidate answer than c2.

The rationale behind SCO-QAT is as follows. We use the retrieved passages as a resource to look up question terms and locate the correct answer. When a set of question terms QT co-occurs with an answer A, it indicates that some kind of relation exists between the QT set and the answer A, which could be helpful for identifying correct answers. However, since this type of relation is not always correct, we have to find a way to deal with noisy relations. We use the confidence score given in Equation (1) to measure the goodness of a rule, which is similar to the method used for finding association rules. Then, we sum the confidence scores of all the co-occurrences of all question term combinations to resolve the noisy rule problem. This technique is useful if there is a lot of redundant information about a given question and answer in the returned passages.

4. System Performance

In terms of RU-accuracy, ASQA2 achieved 0.553 in the C-C subtask and 0.34 in the E-C subtask. They were the best performances for both subtasks at NTCIR-6. When the performance was measured in terms of the R-accuracy, the accuracy decreased to 0.52 for the C-C subtask, which was still the best accuracy rate, and to 0.253 for the E-C subtask, which was the second highest accuracy rate.

Table 2: ASQA2 - overall performance

Accuracy	CC Subtask		EC Subtask	
	ASQA2	best	ASQA2	best
R+U	0.553	0.553	0.34	0.34
R	0.52	0.52	0.253	0.253

In the following sub-sections, we report on some experiments that we conducted to evaluate the properties of our innovations.

4.1. English Question Classification Performance

For question informer prediction, the experimental results show that the proposed hybrid GA-CRF model of question informer prediction improves on the accuracy of the traditional CRF model. By using GA to optimize the selection of the feature subset in CRF-based question informer prediction, we improved the F-score from 88.9% to 93.87%, and reduced the number of features from 105 to 40. Note that the fitness function was used to evaluate the test dataset (UIUC Q500) with the training dataset (UIUC Q5500). In addition, the accuracy of our proposed GA-CRF model for the UIUC dataset was 95.58% compared to 87% for the traditional CRF model reported by Krishnan et al. Thus, the proposed hybrid GA-CRF model for question informer prediction significantly outperforms the traditional CRF model.

For English question classification, the fine-grained accuracy was 82.32% for 10-fold cross validation on the training dataset (IASLEQ5288E), and approximately 88.79% for coarse-grained accuracy. The features used for SVM-based English question classification were WB (word bi-gram), F1 (first word), F2 (first two words), QIF (question informer), QIFB (question informer bi-gram), and WH (question wh-word, 6W1H1O: who, what, when, where, which, why, how, and others).

We also conducted an experiment on the training data of IASLEQ5088E and the test data of CLQA1T200E. The experimental results show that we enhanced the fine-grained accuracy of English Question Classification (EQC) from 68.0% (WB) to 78.5% (WB+F1+F2+WH+QIF+QIFB) by using Support Vector Machines (SVMs). Meanwhile, we increased the coarse-grained accuracy from 71.0% to 83.5%.

We used the 5,288 questions as our training dataset and the WB+F1+F2+WH+QIF+QIFB features to train our SVM model for the test dataset, which was taken from NTCIR-6 CLQA's formal run of 150 English questions. (CLQA2T150E) The experimental results were as follows.

The top 1 accuracy of fine-grained English question classification was 94% for CLQA2T150E. The experimental results of English question classification using different features in SVM models are shown in Figure 1. It is significant that, by integrating GA-CRF-based question informer prediction as a feature, the SVM-based English question classification model performs better than the model that uses the baseline word-based bi-gram feature.

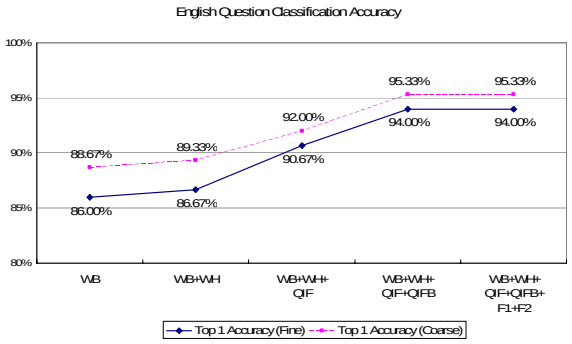

Figure 2. Experimental results for English Question Classification (EQC) using SVM

4.2. Answer Template Filter Performance

Although the Answer Template Filter can only cover some questions, it performed quite well when it was applied. To demonstrate its effectiveness, we only analyzed the questions to which the Answer Template Filter can be applied. Table 4 lists some samples of the 126 Answer Templates that we generated.

For the NTCIR-6 CLQA C-C subtask, the question coverage was 37.3% and the RU-accuracy for the questions covered was 0.911. Meanwhile, for the E-C subtask, the question coverage was 20.7% and the RU-accuracy was 0.807. The accuracy rates were both much higher than the overall accuracy rates, which were 0.553 and 0.34 respectively.

Intuitively, it seems that Answer Template Filters can only deal with short questions. However, we find that, although the average length (14.5 characters) of the covered C-C questions is smaller than the average length (15.27 characters) of all questions, there are still some long questions that can be answered correctly; for example, the question CLQA2-ZH-T3063-00 and the question CLQA2-ZH-T3126-00. In fact, the Answer Template Filter was quite effective in the C-C subtask. When it was applied, 13 more questions were correctly answered and only 2 more questions failed.

Table 3: Answer Template Samples

ARTIFACT Na PERSON VE ，
LOCATION Na PERSON – VC
LOCATION Na P PERSON LOCATION Na – Na
OCCUPATION Na - V - PERSON
ORGANIZATION N - PERSON PA N
PERSON 表示 ， ORGANIZATION
PERSON 是 ORGANIZATION - OCCUPATION
TIME LOCATION Na - Na
Na PERSON P PERSON Nd V
VH 的 LOCATION OCCUPATION PERSON
最 VH 的 OCCUPATION PERSON

4.3. SCO-QAT Performance

The SCO-QAT answer ranking feature performed well on the CLQA datasets. To observe its effect, we removed the Answer Template filter and conducted some experiments.

The experiments showed that, for NTCIR-5 CLQA C-C and the E-C datasets, using only SCO-QAT to rank answers achieved 0.505 and 0.21 RU-accuracy respectively, which were both higher than the best RU-accuracy for the C-C and E-C subtasks in NTCIR-5. For NTCIR-6 questions, we achieved 0.4 RU-accuracy for C-C and 0.273 RU-accuracy for E-C using only the SCO-QAT feature. When combined with other features and training the weights on the NTCIR-5 dataset with a Genetic Algorithm, the performance improved slightly to 0.46 for C-C and 0.28 for E-C. These experiments demonstrate the effectiveness of SCO-QAT as an answer ranking feature.

However, SCO-QAT failed when dealing with some questions. For example, for the question "誰是海峽兩岸關係協會主席？", there are several lengthy passages [1] containing all the question keywords, but they convey a completely different meaning to that of the question. It is impossible to distinguish the correct answer in such cases with

[1] such as 「美國在台協會理事主席卜睿哲十六日在波士頓近郊的佛萊契外交學院舉行的「台灣關係法」二十週年研討會中表示，美國將注意兩岸關係未來數月的發展，希望海峽兩岸能發展出建設性、實質性的模式以降低緊張」

co-occurrence-based methods like SCO-QAT. In addition, some questions failed due to synonym problems. When the answer of the same question was mentioned in the CIRB20 corpus, all the occurrences used "會長", instead of the question keyword "主席".

5. Discussion and Conclusion

For NTCIR6 CLQA, we built an English Question Classification sub-module and two advanced shallow techniques, Answer Templates and SCO-QAT, to deal with both C-C and E-C subtasks. We achieved 0.553 RU-accuracy in the C-C subtask and 0.34 RU-accuracy in the E-C subtask.

Although low coverage techniques, such as the Answer Template Filter, could only deal with some questions, we found that they could still be useful if the accuracy was high enough when the technique applied. Although the coverage of the Answer Templates was below 50%, it helped boost our QA performance substantially. The key point to consider when incorporating such a low coverage technique is to identify situations in which it can work.

We also found that global information obtained from all the returned passages is very useful in a QA system In NTCIR-5; we only considered local information obtained from a single passage. However, in NTCIR-6, both the Answer Template Filter and the SCO-QAT feature considered all the occurrences of an answer, and both achieved good results. It is not clear whether the results were due to the nature of Chinese-related QA, the corpus used, or the way the questions were created. Further research is needed in this area.

In the E-C subtask, we were surprised that the Answer Template Filter and the SCO-QAT feature were not influenced much by the noise introduced by machine translation. We think that, because they do not consider the syntax of a question, they can perform well in both mono-language and cross-language situations.

Finally, to facilitate better QA research and more reliable evaluation, we suggest that the number of test questions should be increased[2]. It would also be useful to have some kind of version control service to allow researchers to add new answers for the standard questions to support post-hoc experiments.

6. Acknowledgments

This research was supported in part by the National Science Council under Grant No. NSC94-2752-E-001-001-PAE.

We would like to thank the Chinese Knowledge and Information Processing group (CKIP) in Academia Sinica for providing us with AutoTag for Chinese word segmentation.

7. References

[1] Cross Language Evaluation Forum (CLEF), http://www.clef-campaign.org/

[2] Google Translate, http://www.google.com/translate_t

[3] NTCIR Workshop, http://research.nii.ac.jp/ntcir/

[4] Text REtrieval Conference (TREC), http://trec.nist.gov/

[5] M. Y. Day, C. H. Lu, C. S. Ong, S. H. Wu and W. L. Hsu, Integrating Genetic Algorithms with Conditional Random Fields to Enhance Question Informer Prediction, *Proceedings of the IEEE International Conference on Information Reuse and Integration (IEEE IRI 2006)*, Waikoloa, Hawaii, USA, 2006, pp. 414-419.

[6] V. Krishnan, S. Das and S. Chakrabarti, Enhanced Answer Type Inference from Questions using Sequential Models, *Proceedings of HLT/EMNLP* (2005), pp. 315–322.

[7] C.-W. Lee, C.-W. Shih, M.-Y. Day, T.-H. Tsai, T.-J. Jiang, C.-W. Wu, C.-L. Sung, Y.-R. Chen, S.-H. Wu and W.-L. Hsu, ASQA: Academia Sinica Question Answering System for NTCIR-5 CLQA, *NTCIR*, 2005.

[8] D. Ravichandran and E. Hovy, Learning surface text patterns for a Question Answering system, *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 2001, pp. 41-47.

[9] M. M. Soubbotin and S. M. Soubbotin, Use of Patterns for Detection of Likely Answer Strings: A Systematic Approach, *TREC-2002*, 2002.

[2] It is usually more than 400 questions in TREC QA Track.