NTCIR-5 English-Chinese Cross Language Question-Answering Experiments using PIRCS

Kui-Lam Kwok, Peter Deng, Norbert Dinstl and Sora Choi

Computer Science Department, Queens College,
City University of New York, Flushing, NY 11367, USA
kwok@ir.cs.qc.edu, peterqc@yahoo.com, emc21@earthlink.net, sorac@hotmail.com

Abstract

We participated in the English-Chinese CLQA task with the following procedures. An English question was first classified as to its answer category, and then rendered into Chinese in three ways: raw text translation by MT, extracted entity translation by our web-translation algorithm, and web-assisted question expansion followed by MT and entity web-translation. A combined Chinese question is formed that retrieves the top 100 sentences from the target collection. Candidate Chinese entities are extracted from the sentences and ranked for answer-hood based on a combination of five sources of evidence: category, frequency, proximity, web, and similarity.

Results show that when only top-1 answers are considered, 25 questions are answered correctly out of 200 with supporting documents, giving an accuracy and MRR of 0.125. When unsupported answers are included, these measures improve to 0.165. If top-5 answers are considered, accuracy and MRR attain values of 0.325 and 0.1968. When unsupported answers are also included, these measures improve to 0.415 and 0.257.

Keywords: *English-Chinese Cross language Question Answering; web assistance; web-translation of entities.*

1 Introduction

Question-answering (QA) attempts to extract exact answers from documents to satisfy a question. [1]. Cross language QA (CLQA) additionally allows users to pose questions in a source language different from the target document language. The language pair we studied was English-Chinese. Previous investigations have studied CLQA among European languages in CLEF (e.g. see participants in [2]), and between Hindu and English [3]. NTCIR-5 environment involves 200 questions and a large Chinese collection whose characteristics are described in the Overview paper of this proceeding [4]. The questions are limited to factoid questions only. There are many approaches to QA varying from more language understanding oriented to highly statistical oriented. Our approach is in the latter category. This paper is organized as follows: Section 2 describes the process for our CLQA experiments. Section 3 documents the resources we employed. Section 4 describes our method of ranking the candidates and selecting an answer. Section 5 summarizes our experimental results. Section 6 has our conclusions.

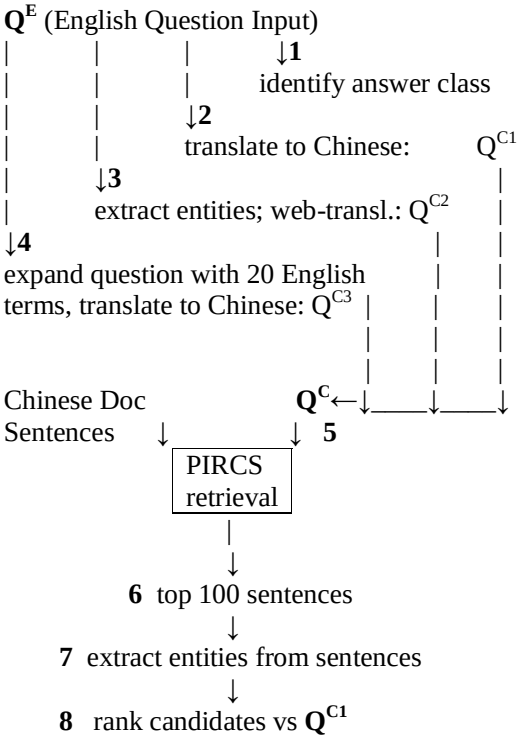

Fig.1 CLQA Processes

2 English-Chinese CLQA Processes

We approach the CLQA problem with the flow diagram shown in Fig.1. Since it is easier to analyze English language than Chinese and there are more tools for it also, we perform as much processing as possible with a given source English question. In Fig.1, four operations were done when an English

question Q^E is received: (1) determine its answer class – e.g. does the question require a person or numeric for its answer? (2) translate the raw English question to Chinese; (3) extract entity names from the question and translate these to Chinese using our web-translation procedure – this increases the chance of getting correct translations to entity names or terminology present in the question; (4) expand the English question by twenty terms via web searching and then translate to Chinese – the web is regarded as an all-domain thesaurus, and this helps to enrich the question with more related words.

A Chinese query Q^C was then constructed by concatenating all of the above three translations (5). This query is used to do sentence retrieval from the target Chinese collection via our PIRCS engine, and the 100 top sentences are returned (6). Entities are extracted with their class tags from these sentences. The entities are then ranked based on five sources of evidence to be discussed in Sec.4. The top ranked candidate becomes our answer to the question. If extraction does not result in any candidate from the top 100 sentences, we assume no answer exists. This did not occur in these experiments.

3 Resources for English-Chinese CLQA

As alluded to in Sec.2, a number of resources were employed to complete the CLQA task. Some of the resources we built in-house, and others we rely on external sources. These are discussed in the following subsections.

3.1 Question Classification

An important issue in QA is to discover what a question wants. Initially, we intended to discover more refined question classes based on the categories provided by the Cognitive Computation Group at the University of Illinois, Urbana-Champaign (http://l2r.cs.uiuc.edu/~cogcomp/Data/) which has a total of 50 categories. This was done by POS tagging a question using MXPOST (http://www.cogsci.ed.ac.uk/~jamesc/taggers/MXPOST.html), followed with parsing by Collins parser (http://people.csail.mit.edu/mcollins/code.html). Our program detects keywords like 'Who', 'When', 'Where', 'Which', 'What', 'How' in a question, and the noun phrase(s) following it. Based on the words in the nearest noun phrase, relevant entity word lists and simple heuristics, we detect the possible answer class to an English question. Later, this may provide better evidence that a selected candidate from retrieved sentences satisfies answer-hood. Since NTCIR-5 CLQA task focuses on six categories (person, location, organization, date/time, number, artifact), we mapped the refined classes to the first five plus an 'unknown' class that catches everything else including 'artifact' which we did not explicitly attempt to identify.

We tested this classification procedure on the training questions (T0001-T0200), and it provides an accuracy of about 80%. We believe this is much more accurate than the other processes to be discussed.

3.2 MT Software

One of the most important steps for E-C CLQA is translation from an English question to Chinese that can retain the fidelity and intent of the question. As in our CLIR attempts, we employ Systran MT (http://www.systransoft.com/index.html) as the basis (Steps 2,4) since it can provide reasonable translation of common English. Systran Chinese output can be segmented or not. We used the un-segmented output at Step 2 and the segmented output at Step 4 to add variety to the final Chinese question Q^C for retrieval purposes. As examples, we show the translation output for questions T1017 and T1057:

T1017: *Which Chinese singer is the first Asian student of Pavarotti of The Three Tenors?*

Systran: *哪位中國歌手是三個進程的帕□洛蒂的第一亞裔學生?*

T1057: *Who is the first Berlin mayor to publicly admit that he is a gay?*

Systran: *誰是第一柏林市長公開地承認,他是 gay?*

We notice that for T1017, one character of the transliteration of 'Pavarotti' (□) is missing, and that 'Tenors' also has wrong translation (进程). For T1057, 'gay' failed to get translated by Systran.

3.3 Web-Assisted Entity Translation

Systran may not translate named entities well, especially current ones. We augment its output with our entity-oriented online translation software CHINET [5] at (Steps 3, 4). CHINET combines web-based translation, and special transliteration procedures for Chinese location and person names expressed in Pinyin (English). Web-based translation employs text patterns that may occur in Chinese web snippets returned after web querying with an English named entity or terminology. The English named entity is extracted from a question using BBN's IdentiFinder (Sec. 3.3). Web translation is ideal for current and popular named entities or terminology. It is always up-to-date and has good accuracy when patterns are found. This complements well with Systran which is a static package.

For example, question T1017 'Pavarotti' was correctly extracted as an entity, and our web-assisted translation gives five alternative results. For T1057, 'Berlin' was also extracted and translated correctly via web-based translation. These are shown as follows:

T1017:

(((PAVAROTTI))) *巴伐洛堤*
^ 帕華洛帝
^ 帕瓦羅蒂
^ 帕瓦洛帝
^ 帕瓦羅帝

T1057:

(((BERLIN)))柏林

3.4 Named-Entity Extraction

Our entity translation (Sec.3.3) works well when an isolated named entity is given as input. Ordinary English words or phrases lead to diverse content in the snippets that are returned. The translation may sometimes work but results are unpredictable. Thus, we employ an entity extraction software provided by BBN called IdentiFinder [6]. It is based on an HMM decoder and has facility to extract entities from both English questions (Step 3) and Chinese document sentences (Step 7). The software not only identifies entity objects, but also tags them as to what category each object belongs to. IdentiFinder can detect five of the NTCIR-5 six categories except for 'artifact'. Section 3.3 showed examples of entities extracted from the English questions.

3.5 Web-Assisted Query Expansion

If one does *direct* QA from a collection, then one needs to retrieve the relevant documents, passages or sentences that may contain an answer to the question. However, a user question may not be well-composed, or it may be short and has insufficient text variety for a good retrieval. To overcome this, we perform question expansion before retrieval by using our web-assisted English-English query expansion software [7]. This software assumes that the web is an all-domain thesaurus, and a search engine (Google) was employed to return relevant snippets based on an input question. Sometimes, a question is long, and probing the web with such long questions may return null pages. We employed a 'window rotation' method [7] to break a long question into short overlapping ones. Final retrieval list is defined by a voting process from the multiple lists, and this defines expansion terms for the input question. We have employed an expansion of 20 terms.

An interesting observation is that sometimes the returned expansion terms may actually contain an answer to the question (in English). Some investigators previously performed *indirect* QA (e.g. [8]) successfully by finding an answer from the web first, and then locate supporting documents for that answer from the target collection. This however may not be as successful in a CLQA environment because there may be less chance of obtaining an answer in English pages when the question content is Chinese information oriented. Moreover, a translation of a candidate answer needs to be performed, and this adds further uncertainty. We instead translated all the expanded terms (including potential answers) as question terms for enhancing the Chinese retrieval process. Later, the translated expansion terms are also used to confer evidence of answer-hood for candidates as discussed in Sec.4.

As examples, the expansion terms derived for T1017 and T1057 via web-assistance are shown below. Their translations later via both Systran and web-translation are also given.

T1017:

Expansion Terms: *pavarotti, tenors, three, luciano, chinese, concert, domingo, opera, singer, music, carreras, asian, placido, tenor, china, beijing, jose, student, singers*

Systran: *帕□洛蒂 進程 三位 luciano 中國 音樂會 多明哥 歌劇 歌手 音樂 carreras 亞裔 placido 進程 瓷 北京 jose 學生 歌手*

Web-based Translation:

(((LUCIANO))) *盧西安諾*
^ 露西亞諾
^ 陸西阿諾
(((CARRERAS))) *卡雷拉斯*
^ 卡列拉斯
^ 卡瑞拉斯
(((PLACIDO))) *普拉西多*
^ 普拉契多
^ 普拉希多
(((JOSE))) *何塞*

T1057:

Expansion Terms: *gay mayor admit publicly berlin marriage lesbian index who history archive rights London opinion wowereit klaus report crb most*

Systran: *快樂 市長 承認 公開地 的柏林 婚姻 女同性戀的 索引 歷史 檔案 糾正 倫敦 觀點 wowereit klaus 報告 crb 最*

Web-based Translation:

(((WOWEREIT))) *萊特*
(((KLAUS))) *克勞斯*

The expansion terms added reasonable context to both question. For T1017, the focus may have shifted to the three tenors rather than a student of Pavarotti. It also does not contain the answer. It can also be seen that our web-translation was successful in transliterating the modern opera singer's first or last names while Systran failed.

3.6 Sentence Retrieval

In QA, it is quite likely that answers are in close proximity to the related keywords in a question. Since Chinese sentences in general have dense information, we decided to compose each sentence as a sub-document for retrieval with our PIRCS system

(Step 6). PIRCS uses a probabilistic retrieval algorithm that departs from the usual like Okapi by making use of collection frequency of a term instead of document frequency, and one of its combined retrieval formulae can be reduced to one like that of a simple language model [12]. Sub-document segmentation (having size of hundreds of Chinese characters) is routinely done in our PIRCS. However, this is the first endeavor to do retrieval using sub-documents of the sentence granularity.

Example sentences retrieved are given below for the two questions under study. They have been processed by IdentiFinder for entity extraction.

T1017:

mhn_xxx_20010725_101393901 20010725
記者<ENAMEX TYPE="PERSON">黑中亮</ENAMEX>／綜合報導<ENAMEX TYPE="PERSON">歌王帕華洛帝</ENAMEX>一生以演唱爲事業，極少有收學生的打算，但年齡爲三大男高音之首的他，近日終於答應決定在<ENAMEX TYPE="LOCATION">北京</ENAMEX>收中國歌唱家<ENAMEX TYPE="PERSON">戴玉強</ENAMEX>爲徒，成爲<ENAMEX TYPE="PERSON">帕華洛帝</ENAMEX>亞洲第一位學生外，並將於<TIMEX TYPE="DATE">今年 12 月初</TIMEX>再來<ENAMEX TYPE="LOCATION">中國</ENAMEX>，更是他頭一回在<ENAMEX TYPE="LOCATION">上海</ENAMEX>演唱。．

T1057:

mhn_xxx_20010806_10321190> 20010806．
自<TIMEX TYPE="DATE">六月初</TIMEX>，現任<ENAMEX TYPE="LOCATION">柏林</ENAMEX>市長<ENAMEX TYPE="PERSON">渥維雷特</ENAMEX>在競選市長期間，亮出「我是同性戀，這也滿好的」的口號，同性戀在<ENAMEX TYPE="LOCATION">德國</ENAMEX>便不再是禁忌話題。．

It is seen that in the sentence for Question T1017, a person entity candidate (***戴玉強***) is actually the answer. Unfortunately, answer selection (Sec.4) ranks another one to the top and our system failed for this question. For T1057, a candidate in this sentence is also the answer and our answer selection succeeded for this question.

4 Answer Ranking and Selection

The processes for our CLQA experiments are described in Fig.1 and explained in Sections 2 and 3 for Steps up to (6). After the top 100 sentences have been retrieved for a question, one needs to identify a unique candidate Chinese string as an answer to the question, or rank for the top 5.

We pass each sentence through BBN's IdentiFinder (Chinese version) which extracts and tags possible entity strings with their categories (see examples in Sec. 3.6). These candidates are captured in a table for evaluation, together with their source (DocID and sentence#) and their rank position during retrieval. For each candidate, five intuitive measures were evaluated as evidence of its potential as an answer.

(a) Categorical Evidence

At Step (1) of Fig.1, the answer category of a question Q^E has been evaluated. These classes have either been assigned to the five provided by IdentiFinder, or collapsed into an unknown class. Since the candidates from the retrieved sentences are also tagged, we used their category agreement as one source of evidence that a candidate is indeed an answer. Since both procedures (our question classification and IdentiFinder) have uncertainty, we used a graded measure Vc for this agreement:

```
if categories agree        { Vc = 10 }
else {
   if (both are tagged as person or location or
organization)      { Vc = 5 }
   else            { Vc = 1}
}
```

When the categories between the question and a candidate sentence entity do not agree and both are named entities, we give them a medium matching score (Vc=5) since the process sometimes have tendencies to mix up the named entities, such as identifying a location as organization or an organization as a person. However, the likelihood of mistaking named entities with other types such as percent or date is low, and Vc is scored as the default value of 1 (no value) when there is mismatch.

(b) Frequency Evidence

Each possible candidate may appear in different sentences. Its sentence occurrence frequency f is also captured. We assume that, based on repeated confirmation, the more often a candidate occurs, the more likely that it is a correct answer. We use the following measure Vf to capture this information:

$$Vf = 1 + 0.1 * \log (f)$$

(c) Web Evidence

At Step 4 Fig.1, an English question was expanded with 20 terms from the web. These 20 terms not only enrich the question context, but they might contain the answer (in English) to the question as well. In previous monolingual work, investigators do indirect QA by extracting an answer from these web pages (or other external sources), and later identify a document or sentence that contains this answer. Here, we do not attempt to extract answers from the web pages. However, every candidate (extracted from the retrieved sentences) is compared with the expanded terms and separately with the original translated query $Q^{C1} \cup Q^{C2}$, and assign a

value Vw as follows:

```
Vw =1; /*default value*/
if (candidate occurs in translated original question) { Vw =0; }
else if (candidate occurs in translated expansion terms)
        { Vw =2; }
```

We assume that if a candidate occurs in the set of expanded terms, it has a higher chance of being an answer, but that an answer should not appear in an original (translated) question statement.

(d) Proximity Evidence

When a candidate is detected in a sentence, we assume that the closer this candidate is to some of the question keywords (that appear in the same sentence), the more probable it is an answer. Proximity measures are accumulated both preceding and succeeding a candidate. Suppose multiple question words are found in a sentence containing one or more candidates. For each candidate, evaluate a preceding score Vp-pre and a succeeding score Vp-suc, and the final proximity score Vp is the sum of both.

Evaluate Vp-pre score:

```
for (each c preceding a candidate in a document sentence)
{
        while (c is a substring of the translated question){
        Vp-pre+=match-length/
                        (distance-from-candidate)^2 ;
        c=c || next_c;
        }
}
```

Here, c is generalized to a Chinese character, numeric sequence, or English word). This means a long sequence of word matching will be given high weights including its subsequences. A similar procedure for evaluating Vp-suc is also done for patterns appearing after a candidate.

(e) Similarity Evidence

If a sentence has high similarity to the original translated question, its candidates may also be more likely to be possible answers. Thus, a similarity value Vs is calculated for each retrieved sentence with respect to the given question. In determining this score, the rank of a sentence in the retrieval list is also used. Let m_i (i = 1..5) be the number of word matches of length i between a sentence and a question. Then,

$$Vs = (\sum_{i=1..5} m_i * sqrt(i))/sqrt(sentence\text{-}length)/rank^{0.25}$$

Vs gives higher value to longer matching words normalized by the sentence length. We limit the longest matching length to 5.

There are uncertainties in every step of our processes. For example, we do not have 'artifact' as an answer category and 'unknown' is our catch-all assignment when classification failed. Translation can fail, web expansion terms can be noisy, and similarity calculation may be unreliable. Combination of these five different sources of evidence (by multiplication) may lead to a more robust single score for determining answer-hood of a candidate. All component scores have default value of 1, so that their absence means they have no influence in the final score. We made three submissions differing only in the answer selection strategy: pircs-E-C-01 uses all Vc*Vf*Vw*Vp*Vs, pircs-E-C-02 uses Vc*Vf*Vw*Vp, and pircs-E-C-03 uses Vc*Vw*Vp.

5 English-Chinese CLQA Results

Results of our E-C CLQA experiments appear in Table 1. Table 1a tabulates results of our three runs when only the top 1 answer is considered. pircs-E-C-01, which makes use of all the evidence for answer selection, returns the better result of 25 correct out of 200 (accuracy 0.125). Number of unsupported answers in top 1 position (8) is also better than the other runs. If unsupported answers are also counted as correct, the accuracy becomes 33/200 or 0.165. The three runs are very close to each other, with pircs-E-C-03 performing better than pircs-E-C-02. The best monolingual CC-CLQA result has Top 1 accuracy 0.375, and 0.445 when unsupported answers are also counted correct. Our best CLQA run is about 1/3 of the monolingual accuracy. Translation inaccuracy probably accounts for a large portion of this deficiency.

Table 1b tabulates results when the Top 5 answers are considered for evaluation. Here, pircs-E-C-(i) i=1 to 3, and pircs-E-C-u-(n+3) are pairs differing only in whether Top-2 to Top-5 answers were returned or not. Counting the additional 4 answer sets in the Top 2 to Top 5 positions give pircs-E-C-u-04 an extra 40 correct answers and an extra 10 unsupported. This leads to a Top 5 accuracy of 65/200 = .325, and

RunID→ pircs-	**E-C-01**	**E-C-02**	**E-C-03**
Right[1]	25	23	24
Unsupported[1]	8	5	6
Accuracy	.125	.115	.12
MRR	.125	.115	.12
Top5	.125	.115	.12
Accuracy+U	.165	.14	.15
MRR+U	.165	.14	.15
Top5+U	.165	.14	.15

(a) Only Top 1 Answer

runs→ pircs-	E-C-u-04	E-C-u-05	E-C-u-06
Right[1]	25	23	24
Right[2]	14	15	13
Right[3]	14	14	12
Right[4]	6	6	4
Right[5]	6	8	9
Unsupported[1]	8	5	6
Unsupported[1]	6	3	3
Unsupported[1]	1	1	2
Unsupported[1]	2	4	1
Unsupported[1]	1	1	1
Accuracy	.125	.115	.12
MRR	.1968	.1913	.1865
Top5	.325	.33	.31
Accuracy+U	.165	.14	.15
MRR+U	.257	.2315	.2296
Top5+U	.415	.4	.375

(b) Includes Top 1-5 Answers

Table 1a,b: Results of CLQA Runs

Top5+U accuracy of 83/200 = 0.415. When Top 5 answers are considered, pircs-E-C-u-05 has a slight edge over pircs-E-C-u-06: e.g. MRR for pircs-E-C-u-05 is .1913 vs .1865 or pircs-E-C-u-06.

According to results, the 200 questions are distributed as: person 80, location 53, organization18, number 16, time 20, artifact 13. With Top 1, our E-C-01 has 25 correct answer distributed as shown in Table 2. Thus, the best recall values are organization (22%), location (17%), time (15%) and person (8.8%) categories. When one accumulates the Top 2 to Top 5 entries, taking the precaution of removing duplicate answers that have been seen in earlier ranks, location and organization categories still have the best recall values followed by time and person. At Top 1-5, recall of person category however is more than 2.5 times of that at Top 1 rank.

	per	loc	org	num	time	artif
Given	80	53	18	16	20	13
Top 1	7	9	4	1	3	1
%recall	*8.8*	*17*	*22*	*6.3*	*15*	*7.7*
Top1-2	13	12	5	2	4	1
%recall	*16*	*23*	*28*	*13*	*20*	*7.7*
Top 1-3	16	17	5	2	5	2
%recall	*20*	*32*	*28*	*13*	*25*	*15*
Top 1-4	17	18	6	2	5	2
%recall	*21*	*34*	*33*	*13*	*25*	*15*
Top 1-5	18	19	6	2	6	2
%recall	*23*	*36*	*33*	*13*	*30*	*15*

Table 2: pircs-E-C-01 Recall Results

6 Conclusion

Our experiments with 200 questions showed that it is possible to provide English-Chinese CLQA capability using the available tools. However, result based on using Top 1 answers only is low (accuracy 0.125). Top 1 poses a very strict condition. If one relaxes the restrictions to Top 5, accuracy improves to 0.325. Just as in CLIR, translation accuracy probably accounts for much of the deficiency.

Acknowledgments

We thank BBN for providing the IdentiFinder entity extraction software used in this experiment.

References

[1] Voorhees, E.M & Tice, D.M. The TREC-8 question answering track evaluation. In: Information Technology: The Eighth Text REtrieval Conference (TREC-8). NIST Special Publication 500-246. pp.83-105, 2000.

[2] Magnini, B, Romagnoli, S, Vallin, A, Herrera, J, Peñas, A, Peinado, V, Verdejo, F & de Rijke, M. The multiple Language Question Answering Track at CLEF 2003. Available at: http://clef.iei.pi.cnr.it/

[3] Sekine, S and Grishman, R. Hindi-English cross-lingual question-answering system. ACM TALIP 2(3):181-192, September 2004.

[4] NTCIR-5 CLQA Overview paper, this volume.

[5] Kwok, K.L, Deng, P, Sun, H.L, Xu, W, Dinstl, N, Peng, P. & Doyon, J. CHINET – a Chinese name finder for document triage. Proc. of 2005 International Conference on Intelligence Analysis. Available at: https://analysis.mitre.org/ proceedings_agenda.htm#papers.

[6] Bikel, D.M, Miller, S, Schwartz, R & Weischedel, R. A high-performance learning name-finder. In: Proc. on Conference of Applied Natural Language Processing, 1997.

[7] Kwok, K.L, Grunfeld, L, Sun, H.L & Deng, P. TREC2004 robust track experiments using PIRCS. In: Information Technology: The Fourteenth Text REtrieval Conference TREC-2004. Available at: http://trec.nist.gov/pubs/trec13/papers/queens-college. robust.pdf

[8] Brill, E, Lin, J, Banko, M, Dumais, S & Ng, A. Data-intensive question answering. In: Information Technology: The Tenth Text REtrieval Conference, TREC 2001. NIST Special Publication 500-250. pp.393-400, 2002.

[9] K.L. Kwok. Improving English & Chinese Ad-Hoc Retrieval: A Tipster Text Phase 3 Project Report. Information Retrieval, 3:313-338, 2000.