Component Analysis of a Chinese Factoid Question-Answering System

Kui-Lam Kwok
Computer Science Dept., Queens College, City University of New York
kwok@ir.cs.qc.cuny.edu

Abstract

An analysis is provided for three major components of a simple Chinese Question-Answering system: passage retrieval, entity extraction and candidate selection. The order of least effective component is determined to be: answer selection, retrieval and extraction. In cross-lingual QA, deficiencies in question translation not only lead to retrieval loss, but may also have adverse effects at answer selection.

Keywords: *Passage retrieval; entity extraction; answer selection; translation effect on answer selection.*

1. Introduction

A simple approach to monolingual factoid question-answering (QA) is shown in the upper part of Fig.1 consisting of question analysis, passage retrieval, entity extraction and answer selection. A question is first analyzed to discover what entity type it needs as answer (such as person, location, etc.). Passage IR attempts to use a (modified) question to isolate a small subset of good passages from the target collection so as to have a reasonable chance of including one or more correct answers (answer-bearing passages). Extraction step processes the retrieved passages to identify all potential entities and their type and form a candidate answer pool. Answer selection screens the candidate pool to rank/select one candidate and its supporting passage as answer to the question. In this paper we focus on the last three steps which involve passing retrieved items from one processing component to another. It seems useful to analyze the quality of the data passed which reflect on the individual as well as relative effectiveness of each component. If multiple systems using this approach are analyzed in this fashion, one may select the best performing module for each component and potentially compose a QA system with better accuracy. We employ a simple 'presence of answer-bearing sentence per question' as quality measure. Most investigations report only the final accuracy of their QA systems, or on effects of particular procedures on the final QA accuracy (e.g. [1,2]). In [3], the paper provides recall values of the retrieval component.

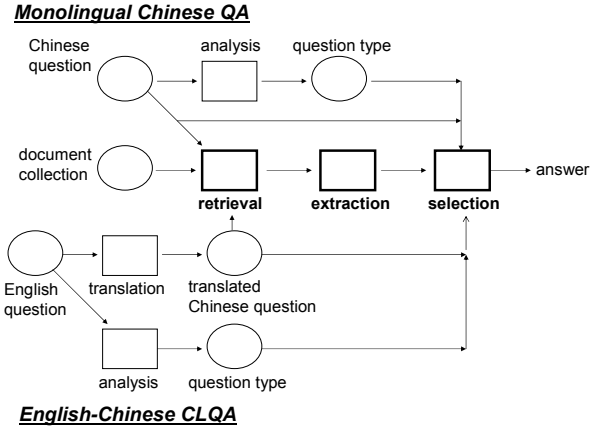

Fig.1 Three major components of a QA System

When QA is performed in a cross-lingual environment (shown in the lower part of Fig.1), the questions are given in a source language different from the target collection language. A popular approach is to translate questions to the target language, which are then used as in monolingual QA. It is well-known that deficiencies of translation can adversely affect cross-lingual IR effectiveness. This paper shows that they may also adversely influence the answer selection component of a QA system compared to the monolingual environment.

In NTCIR-6 & 5, we implemented a system with the above steps for the CLQA task [4,5]. For retrieval, we used our PIRCS retrieval engine with sentence unit as passage. Entity extraction is done using a commercial software package IdentiFinder [6] from BBN. This can extract and type-identify entities for most of the required NTCIR types except for 'artifacts' and Chinese numeric expressions. These latter are extracted with our own in-house developed modules. For answer ranking and selection, we employed a formula that makes use of five statistical evidences for each candidate and its associated sentence. This paper may be considered as additional analysis of our QA work in NTCIR-6 and 5.

2. Component analysis for NTCIR-6 CLQA

NTCIR-6 Chinese language CLQA task consists of 150 questions in both Chinese and English, a collection of nearly ¼ million Chinese newspaper documents, and a set of judged correct answers for the questions [7]. The English questions are assummed to be correct translations of their Chinese counterpart. A QA system

is to return one answer with a supporting document for each question. Table 1 shows results of representative runs from our system for monolingual and English-Chinese cross-lingual QA. For monolingual Chinese, 67 questions have their returned answers and supporting documents correct, leading to an accuracy of .4467. For cross-lingual, only 41 questions have correct answers, leading to accuracy of .2733, a drop of nearly 40%.

Table 1: Representative NTCIR-5, -6 CLQA runs

	# of Questions	Correct Answers	Accuracy
NTCIR-6			
Monolingual	150	67	.4467
Cross-lingual	150	41	.2733
NTCIR-5			
Monolingual	200	65	.325
Cross-lingual	200	31	.155

2.1. Monolingual Chinese QA

The accuracy values in Table 1 show the final result of these QA runs. In Table 2, we break up this QA processing into three separate components, and follow their effectiveness at different stages. Under the `C-C Mono' column `General` section, a summary of the relevant data is tabulated. It is seen that 141 of 150 questions were judged to have answer strings *explicitely* appearing in some sentence(s) in the collection (with 9 questions having none). Since our system is not designed for questions of the later type (e.g. capability to reply `no answer in collection`), we will limit analysis to the 141 questions that have explicit answers. Manual judgment by NTCIR organizers provide a total of 871 unique question-document-entity triplets as `gold standard` answers (q-doc-ans). Since our retrieval unit is a sentence, we like to measure effectiveness with sentences. Each `gold` answer document is decomposed into sentences, and those that explicitely contain one or more answer entities are counted to be 2340, and reduced to 2332 unique question-sentence-answer (q-snt-ans) triplets. When entities are removed from these triplets, they generate 2241 unique question-sentence (q-snt) pairs that contain correct answers.

In the `Retrieval` section under `C-C Mono` column of Table 2, data for the retrieval step are tabulated. This run has used a retrieval depth of 28, leading to 3948 (=141x28) q-snt pairs retrieved. Out of these, only 396 contain correct answers, giving a precision of .1 (=396/3948), and a recall of .177 (=396/2241) in terms of q-snt pairs. These per-sentence effectiveness values are quite low. However, the 396 good sentences are distributed over 119 unique questions, with 22 (=141-119) questions failing to retrieve any good sentence. From a per-question quality point of view, this means that 84% (=119/141) of the questions have at least one good sentence with a correct answer for possible extraction later. This per-question value provides a measure of the data quality to be passed through downstream, and may be viewed as a reflection of the retrieval effectiveness for a QA system.

At the extraction stage, BBN's IdentiFinder [6] software was used as a black box. Given a sentence, it identifies and tags 9 types of entities such as: person, location, organization, percent, etc. This was augmented with our own routines for numeric and `artifact` (such as movie title, etc.) extraction. This produces 18147 unique question-sentence-entity (q-snt-ent) triplets as shown in the `Extraction` section of Table 2. Of these, only 319 triplets overlap with the gold standard set, and they reduce to 312 unique q-snt pairs. There is a loss of 84 (396-312) q-snt pairs which our extraction procedures failed to extract the gold answers. The 312 pairs are spread over 110 unique questions. Thus by itself, extraction stage provides a per-question quality of 92% (110/119), failing in 9 more questions. The per-question data quality after both components (retrieval and extraction) drops to .78 (110/141).

The last component is selection which takes the pool of q-snt-ent candidates from the extraction phase, and rank-select a top candidate as answer. Our rank-selection procedure was described in [4]: this makes use of five types of evidence: agreement of question type with candidate entity type, absense/presence of a candidate entity in the original question, proximity of a candidate entity to question substrings present in a sentence, similarity of a sentence to a question, and the occurrence frequency of a candidate entity in the retrieved sentences. It is a difficult task since the procedure has to select one top answer for each question from among an average of ~120 (18147/150) candidates. Our selection result for NTCIR-6 C-C monolingual is that 67 questions get correct answers, failing for 43. Considering that only 110 questions with one or more correct answers in its pool were passed down, the selection quality by itself is .61 (=67/110). The overall accuracy of the full QA system is therefore .48 (67/141) using questions that have explicit `gold` answers, and .4467 (67/150) when all 150 questions are considered (Table 1).

In summary, ignoring `no-answer` questions and using per-question measure, the weakest component of our monolingual QA system in order is: 1) answer selection having a per-question quality of .61 (Table 2 bold), and affecting accuracy by -30% (from .78 down to .48 (underscored values); 2) sentence retrieval having a quality value of .84, influencing accuracy by -16% at the beginning; and 3) entity extraction having quality of .92 and affecting accuracy by only -6% (.84 to .78). The extraction component behaves quite well. If we had not used our numeric and artifact extraction modules, the effect will be -12%. This type of analysis helps QA system developers focus on improving the weakest link.

2.2. English-Chinese cross-lingual QA

Similar data for the English-Chinese run is also tabulated in Table 2 under the `E-C CLQA` column. This run has the original questions in English, which were rendered into Chinese by our translation procedures using commercial translation software augmented

Table 2: Component analysis: NTCIR CLQA results

	NTCIR-6		NTCIR-5	
	C-C mono	**E-C CLQA**	**C-C mono**	**E-C CLQA**
General				
No. of questions	150		200	
" having sentenses with explicit answers	**141**		**193**	
No. of gold q-doc-ans triplets	871		643	
Unique no. of gold q-snt-ans	2332		2280	
Unique no. of gold q-snt pairs	2241		2201	
Passage Retrieval				
Retrieval depth	28	80	25	80
No. of retrieved q-snt pairs	3948 (28x141)	11280 (80x141)	4825 (25x193)	15440 (80x193)
No. of retrieved gold q-snt pairs	396	451	396	396
q-snt precision	.1 (396/3948)	.04 (451/11280)	.079 (396/4825)	.025 (396/15440)
q-snt recall	.177 (396/2241)	.2 (451/2241)	.18 (396/2201)	.18 (396/2201)
No.of questions w/ gold q-snt	119	107	149	118
Retrieval per-question quality: % of question w/ gold q-snt pair	***.84*** (119/141)	***.76*** (107/141)	***.77*** (149/193)	***.61*** (118/193)
Entity Extraction				
Unique no. of q-snt-ent	18147	55456	21677	68209
No. of gold q-snt-ent	319	349	268	258
No. of gold q-snt	312	339	268	254
No. of questions w/ gold q-snt	110	96	120	92
Extraction per-question quality: % of question w/ gold q-snt pair	**.92** (110/119)	**.90** (96/107)	**.81** (120/149)	**.78** (92/118)
Retrieval+Extraction per-question quality	*.78* (110/141)	*.68* (96/141)	*.62* (120/193)	*.48* (92/193)
Answer Rank/Selection				
No. of questions with correct gold answer	67	41	65	31
Selection per-question quality: % of question w/ answer	**.61** (67/110)	**.43** (41/96)	**.54** (65/120)	**.34** (31/92)
Final CLQA				
CLQA accuracy: (Subset of Questions)	*.48* (67/141)	*.29* (41/141)	*.34* (65/193)	*.16* (31/193)
CLQA accuracy: (All Questions)	.4467 (67/150)	.2733 (41/150)	.325 (65/200)	.155 (31/200)
Rank/Selection: interchange monolingual Chinese with translated English cross-lingual questions				
No. of questions with correct gold answer	51	37	57	31
Selection per-question quality: % of question w/ answer	**.46** (51/110)	**.39** (37/96)	**.48** (57/120)	**.34** (31/92)

with our entity translation via web mining [5]. The translated Chinese questions are then employed as in the monolingual stream. Because of errors in translation, these Chinese questions may contain totally wrong and unrelated wordings, partially correct or approximate translations, ungrammatic string formations, or missing concepts, among some good translated strings. The following shows how these translation inadequacies may affect the component quality.

From past experimentation, it is found that for CLQA, our approach needs to use larger retrieval depths (in the range of 80 to 100 sentences) in order to improve the chance of getting some answer-bearing sentences when using these noisy, inaccurate question translations. For the purpose of comparing a good monolingual run with a good cross-lingual run, our E-C experiment (Fig.2) returns 80 sentences for each question retrieving a total of 11280 (=80x141) q-snt pairs. Out of these, 451 are gold answers spread over 107 unique questions. Thus, the per-question quality is .76. As expected, noisy translated questions lead to an IR result deficit of .08 compared to the monolingual value of .84. Next, extraction produces 55456 q-snt-ent triplets of which only 339 unique q-snt pairs are good, and they are

spread over 96 questions. Within extraction, this produces per-question quality of 90% (96/107), only slightly worse than monolingual's 92% value. After the first two components, data quality now drops to .68. Finally, rank-selection produces correct answers for only 41 questions, leading to a per-question selection quality of .43, much worse than monolingual's .61. The final QA accuracy is .29 if (`no-answer` questions are ignored) and .2733 for all 150 questions.

The least effective components are still in the order of rank-selection (quality .43), retrieval (.76) and extraction (.9). For CLQA, it is expected that deficiencies in question translation cause data quality loss being passed through (e.g. only 96 questions have sentences with correct answers are passed to the selection phase). At selection, one wants to rank and select an entity candidate as answer based on question terms appearing in sentences. The translated question wordings can be erroneous, ambiguous, or missing compared with the original Chinese question string. This leads to errors in ranking and selection of the extracted entities such as using proximity (between a candidate entity and some translated wordings that co-occur in the same retrieved sentence), or similarity (between translated question and retrieved sentences) calculations. It will also have adverse impact on linguistic approaches to answer selection such as based on syntax patterns [8] or semantics [9]. Question type classification is not a major factor as it was done with the original English questions and has comparable accuracy (89%) to Chinese question classification (86%) in our case.

To observe the adverse effect of translated questions on answere selection, we have repeated the monolingual candidate rank and selection phase, but replacing each original Chinese question string with its translated counterpart. Other conditions such as data stream from retrieval and extraction, question classifcation, etc. are retained and unchanged. This would isolate the impact of cross-lingual Chinese questions (translated from English) on answer selection compared to using the original Chinese questions. (This situation would not occur in the CLQA environment because we assume optimal retrieval and extraction data via the original Chinese questions.) The result is tabulated in the last two rows under `C-C Mono` column of Table 2: selection with translated questions decreases to only 51 questions with correct answers compared to the original 67, leading to a selection quality drop from .61 to .46 of -.15, twice as large as for retrieval (-.08).

An example illustrating the above situation can be seen for Question 63 which has the original Chinese string: `黛安娜王妃的死亡車禍事故發生在哪裡？`. The English counterpart is: `Where did Princess Diana's fatal car accident occur?` which the translation software package rendered to: `戴安娜公主的致命車禍何處發生了?`. This string is augmented with some other terms from web mining for CLQA retrieval, but is used as is for the last stage of answer selection. It is seen that although the translation is good and preserves the original meaning well, it has used alternate terms for the word `Princess` (王妃 vs. 公主), `fatal` (死亡 vs. 致命), and partially different characters for the name `Diana` (黛安娜 vs. 戴安娜). During answer selection, the same sentence was top-ranked as support, but for different entities within it as answer. The sentence is from document udn_xxx_19990810_019009 19990810: `首推「聖體節」，把耶穌及其門徒演成同性戀，另一齣則把英國黛安娜王妃的巴黎死亡車禍，搬上舞台。`. It is seen that the two location entities `英國` (England) and `巴黎` (Paris) agree with the question type, but `巴黎` has close proximity to more question strings `黛安娜王妃`, `死亡車禍` of the original question, and is returned as answer for the monolingual QA case, and is correct. For CLQA, because of issues with the translated string, `英國` is returned as answer for CLQA, and is not correct. This illustrates the influence of question translation on answer selection even for a reasonably good output, not to mention a completely wrong translation.

If the original Chinese questions were employed in the cross-lingual selection phase while maintaining the same cross-lingual data stream input, the result is 37 questions with correct answers (shown in last two rows under `E-C CLQA` column of Table 2). This is similar to the 41 obtained in the cross-lingual run but with a small decrease. In this case, the data input is already corrupted by retrieval via the translated questions. Hence employing the original questions for answer selection does not help, and actually decreases perhaps due to incompatibility. Thus, in the CLQA environemnt, question translation inadequacies may lead to adverse effects on both retrieval data quality and answer selection quality.

3. Component analysis of NTCIR-5 CLQA

Corresponding data for the NTCIR-5 CLQA task [10] with 200 questions and over 900,000 newspaper articles are tabulated in Table 1 and Table 2 under the NTCIR-5 columns. The task is much harder compared to NTCIR-6 for all stages: monolingual per-question retrieval quality is only .77 (compared to .84 for NTCIR-6), and cross-lingual value is substantially worse at .61 vs. .76. Entity extraction is also poorer: quality is at ~.8 compared to ~.9 for NTCIR-6. Answer selection quality ranges from ~.3 to ~.5 compared to NTCIR-6 range of ~.4 to ~.6. There is no change in the least effective component order. NTCIR-5 *monolingual* per-question data quality at selection has a level worse than for NTCIR-6 *cross-lingual* (.62 vs. .68) while that for NTCIR-5 cross-lingual task is much worse at .48. When the translated and original questions are interchanged in the selection tasks, there is similar but less pronounced effect vs. NTCIR-6: 57 correct answers vs. 65 (monolingual), and the same 31 (CLQA). Apparently, the low data stream quality dominates the performance in this task.

4. Conclusion and discussion

A quality analysis of a simple Chinese factoid QA system at the component level was presented. The least

effective component is candidate answer selection (which is not unexpected), followed by retrieval and extraction. For CLQA, deficiencies in translated question not only have adverse effect on retrieval, but also have larger effect at the answer selection stage.

If multiple systems using this approach to QA are analyzed in this fashion, one may be able to compose a better QA system by choosing the most effective module at each of the component stage. For this purpose, one should also include the question analysis and the translation components.

One could also employ more sophisticated metrics to measure the quality of a component. For example, the number of answer-bearing sentences in the component's output data list, or earlier positions of these sentences in the list, can be considered higher quality since it could improve effectiveness downstream.

References

[1] E. Brill, S. Dumais and M. Banko. An Analysis of the AskMSR Question-Answering System. In: *Proc. of Conf. on Empirical Methods in Natural Language Processing (EMNLP)*. pp. 257-264, 2002

[2] B. Katz, et.al. External Knowledge Sources for Question Answering. In: *The Fourteenth Text Retrieval Conference (TREC 2005) Proceedings*. 2005

[3] M.W. Bilotti, P. Ogilvie, J. Callan and E. Nyberg. Structured Retrieval for Question Answering. In: *30th Annual International ACM SIGIR Conference.* pp.351-358. 2007

[4] K.L. Kwok, P. Deng, and N. Dinstl. NTCIR-6 Monolingual Chinese and English-Chinese Cross-Lingual Question-Answering Experiments using PIRCS. In: *Proc. of the Fifth NTCIR Workshop Meeting.* pp.209-214. 2006

[5] K.L. Kwok, P. Deng, N. Dinstl and S. Choi. NTCIR-5 English-Chinese Cross Language Question-Answering Experiments using PIRCS. In: *Proc. of the Fifth NTCIR Workshop Meeting.* pp.209-214. 2005

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

[7] Y. Sasaki, C-J Lin, K-H Chen, H-H Chen. Overview of the NTCIR-6 Cross-Lingual Question Answering (CLQA) Task. In: *Proc of NTCIR-6 Workshop Meeting*, 2007

[8] M. Subbotin. Patterns of Potential Answer Expressions as Clues to the Right Answer. In: *The Tenth Text Retrieval Conf. TREC 2001*. NIST S.P. 500-250. pp.293-302, 2001

[9] S. Harabagiu and A. Hickl. Methods for Using Textual Entailment in Open-Domain Question Answering. In: *Proc. of 21st Intl. Conf. on Computational Linguistics.* pp. 905-912, 2006

[10] Y. Sasaki, H-H Chen, K-H, Chen and C-J Lin. Overview of the NTCIR-5 Cross-Lingual Question Answering Task (CLQA1). In: *Proc of the Fifth NTCIR Workshop Meeting.* pp.175-185. 2005