Improving Answer Ranking Using Cohesion between Answer and Keywords

Shu-Jung Lin Min-Shiang Shia Kao-Hung Lin Jiun-Hung Lin
Scott Yu Wen-Hsiang Lu
Department of Computer Science and Information Engineering
National Cheng Kung University, Taiwan, R.O.C.
{shu-jung, foreverdream, thexfile, jhlin, scottyu }@csie.ncku.edu.tw
whlu@mail.ncku.edu.tw

Abstract

At the NTCIR5 CLQA task, we participated in the Chinese-Chinese (CC) and English-Chinese (EC) QA subtasks. Due to some programming errors in our EC QA system, we will focus on the evaluation of our CC QA system in this paper.

We propose a new method to improve answer ranking using the cohesion between answer and keywords. Our experimental results show the effectiveness of this new method. Besides it, we also use POS and IDF (inverse document frequency) information to extract keywords in questions and answers from the relevant passages.

1. Introduction

We developed a cross-language question answering (CLQA) system and participated in the CC and EC QA subtasks at the NTCIR5-CLQA for the first time. But it is a pity that some programming errors in our EC QA system lead to a bad performance, we thus mainly focus on the CC QA task in this paper. However, it is still effective to employ a Web-based system for dealing with translations of unknown terms and keywords in queries [1, 2]. We will simply introduce the process of query translation in our EC QA system in Section 4.

The original method of ranking answers adopted in our system for NTCIR5 CLQA task has an efficiency problem (see Section 5). To improve the efficiency and accuracy, in this paper we propose a new method to appropriately integrate a more complete relationship between answers and keywords, called "answer-keyword cohesion". We will describe this new method in the next section, and discuss how this new method improves the performance of our system in Section 5. Additionally, we introduce how to help identify keywords in questions and answer candidates in relevant passages by using POS and IDF information in Section 3.

2. Answer Ranking Using Answer-Keyword Cohesion

In general, most methods to rank answer candidates usually based on the scores computed by using the distance between answer candidates and keywords [3, 4], or combining the IDF information with distance like Equation (1) [5, 6].

$$S(A,P_i,Q)=\frac{\sum_{K_j\in P_i\wedge K_j\in Q} IDF(K_j)}{\sum_{K_j\in P_i\wedge K_j\in Q} D(A,K_j)} \quad (1)$$

Here, A is one of all answer candidates extracted. P_i indicates the *ith* passage. Q is the question. $S(A,P_i,Q)$ represents A's score in the passage P_i. K_j is a keyword in this question. $IDF(K_j)$ represents the IDF value of the keyword K_j. $D(A,K_j)$ is the distance between the candidate A and the keyword K_j. Equation (1) only use the distance between an answer and keywords occurring in a passage. However, it may cause a problem. If the number of keywords occurring in the passage is small, but their IDF value are high, then it leads to the score of this answer is high. Table 1 shows an example. The question is "請問人類史上第一顆鑽石約4000年前在哪裡出土？" (Where was the first 4000-year old diamond discovered?). The underlined terms are keywords, and the terms within a rectangle are

answer candidates. The correct answer of this question is "印度" (India). We use Equation (1) to compute scores and rank answer candidates. We can see that the first answer candidate "南非" (South Africa) is wrong, but its score is higher than that of the other candidate because the distance between the candidate "南非" and the only one keyword "出土" is very close, and the IDF value of "出土" is high. The second answer candidate "印度" is correct, but its score is lower than that of the candidate "南非" because many keywords appear in this passage and result in a bigger sum of the distances between the candidate "印度" and keywords.

Table 1. An example of answer ranking using Equation (1).

Query	請問人類史上第一顆鑽石約4000年前在哪裡出土？ (Where was the first 4000-year old diamond discovered?)	
Keywords of question	人類史, 4000, 出土, 鑽石, 年前, 第一	
Correct answer	印度(India)	
The first answer candidate	Answer	南非(South Africa)
	Passage	1860年在南非 出土後，就銷聲匿跡的「南非之星」，也是籌辦幹事之一——礦物學家巴利無意中發現的。
The second answer candidate	Answer	印度(India)
	Passage	人類史上第一顆鑽石 4000 年前在印度 出土。

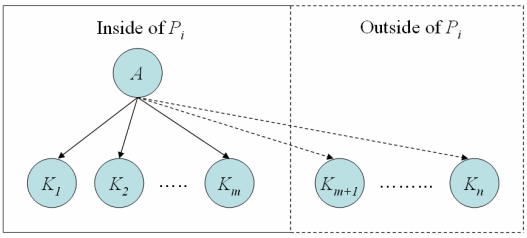

Figure 1. An abstract diagram showing the complete answer-keyword cohesion.

Therefore, we propose a new method to solve this problem by integrating a more complete answer-keyword cohesion (Figure 1). To select the correct answer to a given question, we assume that the best answer should have higher degree of cohesion with all keywords occurring in relevant passages. Also we think that the distance between an answer and keywords outside the passage need to be considered and paid some penalties, which is currently set as the length of the passage. The cohesion score is computed by Equation (2).

$$S(A,P_i,Q)=\frac{\sum_{K_j\in P_i\wedge K_j\in Q} IDF(K_j)+IDF(A)}{\sum_{K_j\in P_i\wedge K_j\in Q} D(A,K_j)+\sum_{K_j\notin P_i\wedge K_j\in Q}|P_i|} \quad (2)$$

where $IDF(A)$ represents the IDF value of the candidate A, and $|P_i|$ is length of the passage P_i. As for the other notations, please refer to Equation (1). We can now get the cohesion score of every candidate in all passages by using the equation (2).

3. Chinese QA System

3.1. Overview

The architecture of our CC QA system is shown in Figure 2. It is separated into three major computing modules: (1) Question Analysis, (2) Document Retrieval and (3) Answer Extraction. The module of Document Retrieval is described in detail in [1] and ignored here, and we will describe the other two modules in the following.

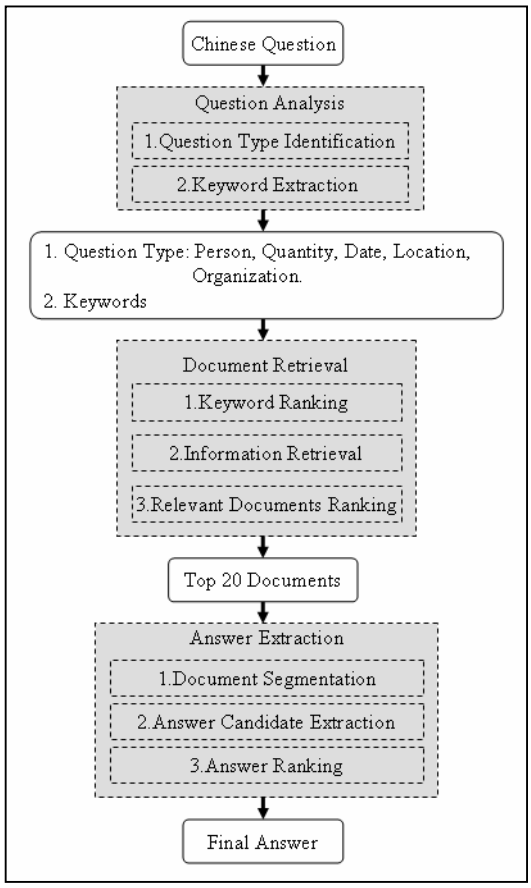

Figure 2. Architecture of our CC QA system.

3.2. Question Analysis

This module is separated into two parts: (1) Question Type Identification and (2) Keyword Extraction.

(1) Question Type Identification

We use rule-based methods to identify question types. For example, the question type of the question "最早發現抗生素的科學家是誰?" can be identified as a person name based on the term "是誰"(the same meaning as "who").

(2) Keyword Extraction

We utilize POS tag and IDF information to extract keywords from questions.

- POS Tagging

We think that noun, verb, number and alphabet term are more important terms in questions, thus we use a on-line representative Chinese tagger (CKIP[1]) to get each term's tag. If the term is tagged as the mentioned POS, then we take them as keywords in questions.

- Keyword Ranking

We use the IDF value to give a keyword candidate a weighted score. Further, we observed that some location names, such as America, Japan, Taiwan, etc., often occur in documents, their IDF value is relatively low. If we use original IDF value to rank these keyword candidates, then some important keywords may be filtered out, and some relevant documents can not be retrieved. Therefore, before extracting final keywords, we increase the weights of some keyword candidates tagged with location names. To save matching time of document retrieval, we use a simple heuristic method to take the first half of keywords as final keywords while the number of keywords is greater than four. Otherwise, we take all keywords as final keywords.

3.3. Answer Extraction

This module consists of three steps: Document Segmentation, Answer Candidate Extraction and Answer Ranking.

(1) Document Segmentation

We think that answer and keywords should often occur in the same passage, so we segment the retrieved relevant documents into passages based on the Chinese period mark "。".

(2) Answer Candidate Extraction

For each question type, we adopt different methods to extract potential answer candidates.

- Person Name

The tag "Nb" in the CKIP tagger indicates proper noun. We submit passages to the CKIP tagger, and then take all terms tagged with "Nb" as answer candidates. To filter out some impossible candidates, we use a simple unigram probability model to estimate the possibility of a person name as Equation (3).

$$P(N) = \prod_{i=1}^{m} P(n_i) \qquad (3)$$

$P(N)$ is the probability of the person name $N = n_1 n_2 \ldots n_m$, $P(n_i)$ means the probability of every Chinese character in the person name N, m is the number of character of N. We show three examples in Figure 3.

Figure 3. Examples showing the estimation of person names by a simple unigram probability model.

We collected two person name lists to train this model. One is the examinee list of Join College Entrance Examination in Taiwan[2], and the other is collected from Chinese documents manually. All of these person names collected contain Chinese, English and Japanese person names.

- Organization Name

Table 2. Suffixes of organization names.

公司(Company)	高峰會(Summit)
集團(Group)	百貨公司(Department Store)
大學(University)	博物館(Museum)
實驗室(Laboratory)	報社(Newspaper Office)
工廠(Factory)	子公司(Subsidiary Company)
銀行(Bank)	中學(Junior/Senior high school)
出版社(Publisher)	藥廠(Pharmaceutical Factory)
樂團(Band)	電腦(Computer Company)
隊(Team)	分公司(Branch Company)
黨(Party)	分行(A branch of company)
組織(Organization)	購物中心(Shopping Center)
政府(Government)	有限公司(Limited company)

We manually identify only 24 suffixes of organization names (Table 2) from 200 test questions, and use these suffixes to learn POS composition rules of organization names from Chinese documents automatically. For example, we use the suffix "藥廠" (pharmaceutical factory) to find "惠氏藥廠" from Chinese documents, and it is tagged as "惠氏(Nb) 藥廠

[1] CKIP tagger: http://ckipsvr.iis.sinica.edu.tw/, developed by Chinese Knowledge and Information Processing Group, Institute of Information Science, Academia Sinica.

[2] http://www.geocities.com/hao510/

(Nc)" by the CKIP tagger. And then we can learn a POS rule of organization name "answer candidate (Nb) + suffix (Nc)" based on the suffix "藥廠". To learn much more POS composition rules of organization names, we will continuously collect more suffixes of organization names in the future.

- Location Name

We manually identify some suffixes of location names and make some POS composition rules to identify location names. Table 3 shows some examples.

Table 3. Some composition rules of location names.

Composition Rule	Example
Nc	台灣(Taiwan)
Nc+Nc	聖地牙哥+機場 (San Diego Airport)
Nb+Nc	中正+紀念堂 (Chang Kai-Shek Memorial Hall)
Nc+Na+Nc	墾丁+國家+公園 (Kenting National Park)
Nc+Nc+Nc	紐約+曼哈坦+公園 (New York Manhattan Park)

- Date and Quantity

To deal with the questions with the question types "Date" and "Quantity", we use the CKIP tagger to extract terms with the tags "Nue" (Numeral), "Neqa" (Number) and "Nd" (Time) as the potential answer candidates.

(3) Answer Ranking

We use Equation (2) to give every answer candidate a cohesion score (Section 2). After getting the score of each answer candidate, we will use two methods to compute the final score for each candidate.

- **Add-all scoring:** Generally, for each question, we may get the same answers from different passages. Thus we can sum up all cohesion scores of these same answers as the final score of the answer as Equation (4).

$$S_{final}(A) = \sum_{i=1}^{n} S(A, P_i, Q) \quad (4)$$

- **Select-best scoring:** It is our observation that, for some cases of questions with the question types "Date" and "Quantity", the add-all scoring method leads to worse accuracy. Therefore, we have the alternative of selecting the best score to be the final score as Equation (5).

$$S_{final}(A) = Max(S(A, P_i, Q)) \quad (5)$$

When we get the final score of all answer candidates by using the (4) and (5), we will choose one of the candidates with the highest score as the final answer.

4. English-Chinese QA System

4.1. Overview

Our EC QA system is separated into four modules: (1) Question Analysis (2) Keyword Translation (3) Document Retrieval (4) Answer Extraction. Because the Document Retrieval and Answer Extraction modules are similar to those of CC QA system, we only describe the Question Analysis and Keyword Translation modulus in the following.

4.2. Question Analysis

We extract two kinds of keywords, including Type-D keywords and Type-U keywords. A Type-D keyword means that the keyword can be translated in the dictionary. A Type-U keyword means that the keyword is an unknown term. We observed that if some English questions contain terms appearing between quote marks or parentheses, then these terms are often proper names. Thus, we extract these terms and take them as Type-U keywords.

Basically, we use an English POS tagger[3] to get tags for each Type-D keywords. Also, we use an English chunker [4] to do phrase identification. For two-word phrases, we take them as Type-U keywords. As for the phrase with more than two words, currently they are difficult to be translated correctly in our system. Thus, we separate them into single words, and take these words as Type-D keywords to be translated by bilingual dictionaries. After the previous processing steps, we can get some important keywords from a question.

4.3. Keyword Translation

To translate keywords in questions, we use two kinds of resources, including a bilingual dictionary and an unknown term translation system. We have developed an unknown term translation system called LiveTrans [5] [2], it utilizes rich anchor texts in the Web and search results from search engines to translate unknown terms. In general, unknown terms and phrases are important terms in the English questions, thus we use LiveTrans to help translate unknown terms, and use a heuristic way to increase IDF value of these important terms. We show an example in Table 4. For the important

[3] http://www.cs.jhu.edu/~brill/

[4] http://pi0657.kub.nl/cgi-bin/tstchunk/demo.pl

[5] LiveTrans: http://wkd.iis.sinica.edu.tw/LiveTrans/lt.html

unknown keyword "Louvre Museum", our system can translate it into "法國羅浮宮" correctly. Actually, LiveTrans can help provide effective translations for some important keywords in questions.

Table 4. An example for effective translation of unknown terms using LiveTrans system.

ID	CLQA1-EN-T1200-00
Question	Which country is the Louvre Museum located in?
Keyword	(Type-D) country (Type-U) Louvre Museum
Keyword Translation	country:國家 Louvre Museum:法國羅浮宮

5. Experimental Results

The evaluation data provided from the NTCIR5 CLQA CC and CE subtasks contains 901,446 Chinese news articles and 200 questions.

5.1. CC Subtask

For the NTCIR5 CLQA task, originally we propose a method to take mutual relations of all keywords into account for enhancing the performance of answer ranking.

$$S(A,P_i,Q)=\frac{\sum_{K_j\in P_i\wedge K_j\in Q} IDF(K_j)+IDF(A)}{\sum_{K_j\in P_i\wedge K_j\in Q} D(A,K_j)+\sum_{K_j\notin P_i\wedge K_j\in Q}|P_i|+\sum_{K_j,K_m\in P_i\wedge K_j,K_m\in Q} D(K_j,K_m)} \quad (6)$$

Table 5 shows the results of our formal runs at the CLQA CC subtask. "R" means that the answers are correct and their corresponding documents can support them, "R+U" means that the answers are correct but the corresponding documents can't support them, and "Number" means the number of the correct answers.

Table 5. Results of our formal runs at NTCIR5 CLQA CC Subtask.

	Number of correct answer	Accuracy
R	64	0.32
R+U	70	0.35

However, we find this method has a efficiency problem in computing distance between all keywords using Equation (6), thus we propose a new method to modify it for reducing computation costs (see Equation (2) in Section 2).

To determine the effectiveness of our new method, we conduct additional runs on the same questions and corpus at the CC subtask to compare performance using three different methods. Table 6 shows that our new method obtains the best performance. Although this new method is only a little better than our original method, it actually saves much computing time.

Table 6. Additional results to compare different methods.

Method	Accuracy
Conventional method (Equation (1))	0.22
New method (Equation (2))	0.36
Original method (Equation (6))	0.35

5.2. Discussion

To improve our new method of ranking answer candidates in the future, we do some error analysis in the following.

(1) Incorrect answers from unrelated passages

Our simple heuristic method to choose the first half of keywords in queries may retrieve unrelated documents or passages, and then extract incorrect answers. In the future, we will investigate how to retrieve relevant documents efficiently and effectively.

(2) Answer ranking error from short passages

In Equation (2), when keywords in questions do not occur in relevant passages, we will use the length of these passages as the distance. As a result, the cohesion score of candidates appearing in short passages may increase improperly.

(3) Tagging error from the CKIP tagger

Our system mainly relies on the CKIP tagger to provide POS information in identifying types of answer candidates, but sometimes some incorrect tags from the tagger lead to incorrect answers. We are considering adding a post-processing step for adjusting the incorrect tags.

6. Conclusion

In this paper, we have proposed a new method to completely utilize the answer-keyword cohesion for improving the accuracy of CC QA system. Although our EC QA system has some problems, it is still effective to translate unknown keywords in questions by employing a Web-based term translation system.

7. References

[1] Jiun-Hung Lin, Min-Shiang Shia, Kao-Hung Lin, Shu-Jung Lin, Scott Yu, Wen-Hsiang Lu. Search-Result-Based Method for Unknown Term Translation in Cross-Language Information Retrieval. *Proceedings of the NTCIR5 Workshop*, 2005.

[2] Pu-Jen Cheng, Jei-Wen Teng, Ruei-Cheng Chen, Jenq-Haur Wang, Wen-Hsiang Lu, Lee-Feng Chien. Translating Unknown Queries with Web Corpora for Cross-Language Information Retrieval. *In Proceedings of ACM SIGIR,* 2004.

[3] Naoya HIDAKA, Fumito MASUI, Keiko TOSAKI. MAIQA:Mie Univ. Participated System at NTCIR4 QAC2. *Proceedings of the NTCIR4 Workshop*, 2004.

[4] Jiangping Chen, He Ge, Yan Wu, Shikun Jiang. UNT at TREC 2004: Question Answering Combining Multiple Evidences. *Proceedings of TREC* 2004.

[5] Masaki Murata, Masao Utiyama, and Hitoshi Isahara. A Question-Answering System Using Unit Estimation and Probabilistic Near-Terms IR. *Proceedings of the NTCIR3 Workshop*.

[6] Dmitri Roussinov, José Antonio Robles-Flores, Yin Ding. Experiments with Web QA System and TREC2004 Questions. *Proceedings of TREC 2004*.