Extracting and Ranking Question-Focused Terms Using the Titles of Wikipedia Articles

Yi-Che Chan　Kuan-Hsi Chen　Wen-Hsiang Lu
Department of Computer Science and Information Engineering
National Cheng Kung University, Taiwan, R.O.C.
{p7694421, p7694413, whlu}@mail.ncku.edu.tw

Abstract

At the NTCIR-6 CLQA (Cross-Language Question Answering) task, we participated in the Chinese-Chinese (C-C) and English-Chinese (E-C) QA (Question Answering) subtasks. Without employing question type classification, we proposed a new resource, Wikipedia, to assist in extracting and ranking Question-Focused terms. We regarded the titles of Wikipedia articles as a multilingual noun-phrase corpus which is useful in QA systems. Experimental results showed that better performance was achieved for questions with type PERSON or LOCATION. Besides, we used an online MT (Machine Translation) system to deal with question translation in our CLQA task.

Keywords: *Question-Focused Terms, Wikipedia, Cross-Language Question Answering*

1. Introduction

Wikipedia[1] is a free Web encyclopedia which is compiled by volunteers from Web. The database of Wikipedia articles contains a large number of specific and proper nouns. It is still growing and may be constructed, edited and corrected by anyone.

Being an article in Wikipedia should follow some rules: popular, easy to understand and being a terminology. Moreover, Wikipedia has over six million articles in 250 languages. This is why we use the title of Wikipedia articles as a multilingual noun-phrase corpus. In the future, more other content in Wikipedia may be utilized as effective resource in related tasks of IR (Information Retrieval).

In this work, we tried to derive question answers by utilizing Question-Focused terms (QF terms) extracted based on this free resource. However, the experience of participation in NTCIR-5 also gave some helps at this work [1]. We will simply introduce the process of how to construct our QA system in Section 4.

Without employing question type classification [2, 3], our experiment results showed that the Top-1 accuracy using our new method was only about 0.153 at the C-C subtask and 0.067 at E-C subtask. After we observed and analyzed the candidates of answers, we found that the accuracy of Top-10 achieved over 0.50 on which questions with type PERSON or LOCATION. However, our system was ineffective for questions with numeric answers. This result may be caused by the insensitivity of dates and numbers both in our QA system and the titles of Wikipedia articles. We will discuss this problem in Section 5.

We used translated results of Google Translate[2], an online MT system, for question translation in our E-C subtask due to its success on terminology translation. The performance of our E-C subtask is nearly one half of that of our C-C subtask.

2. Extracting Question-Focused Terms

Typically, a question focus is a word or phrase in the question that represents the type of the answer [4]. Question focus is often regarded as more informative than question type in QA systems. We defined Question-Focused terms as the keywords in a question, and they should be the most distinguishable words or phrases to extract answers from documents. In our QA system, a QF term is extracted and ranked by its existence in Wikipedia.

There are three steps during extracting QF

[1] Wikipedia: http://wikipedia.org/

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

terms. The steps are described as follows:

Step 1 - Word or Phrase Segmentation: There are no word delimiters (such as spaces in English) in Chinese texts. Therefore, an online Chinese segmentation tool, CKIP[3] tagger, is utilized in our QA system. We used CKIP tagger to divide a question into many segmented words or phrases. These segmented words or phrases are regarded as candidates of QF terms. In addition, CKIP tagger provides POS tag information of each segmented word or phrase. The information can be used in further steps.

Step 2 - Retrieving in Wikipedia: For each segmented word or phrase, we used it as a query to retrieve in the list of article titles of Wikipedia. All exactly matched or partially matched titles are used to compute QF-Score of segmented words or phrases. The method of computing QF-Score will be described in Section 3.

Step 3 – Adding Extra Scores: If the segmented word or phrase exists in the list of article titles of Wikipedia exactly, it may be considered as an important candidate of QF terms and gets extra scores during ranking QF terms. At this step, we also utilized POS tag information to give extra scores for candidates of QF terms which are regarded as nouns.

Table 1. An Example of Question-Focused Term Extraction.

QID	CLQA2-EN-T2003-00
Question	一九九九年時聯合國秘書長是誰？(Who was the UN secretary-general in 1999?)
Candidates of QF terms	**一九九九年(1999)**，**秘書長 (secretary-general)**，**聯合國(UN)**，誰(who)，是(was)，時(in)
Correct answer	安南(Kofi Annan)
Candidate passage #1	安南之前的多數**聯合國秘書長**想必不至於採取他的做法，問題是盧安達、科索伕、東帝汶等地的腥風血雨已經使安南成為一位深具使命感的**聯合國秘書長**。
Candidate passage #2	在**聯合國**方面、**聯合國**官員說，**聯合國秘書長**安南目前正考慮是否前往伊拉克訪問，**聯合國**一個代表團則將於明天前往伊拉克觀察伊拉克當局不准**聯合國**武器檢查員進入的八個地點。

Table 1 shows an example of extracting QF terms. The question is "一九九九年時聯合國秘書長是誰？" (Who was the UN secretary -general in 1999?). After applying Chinese segmentation tool, we can get six segmented words or phrases: 一九九九年(1999), 秘書長 (secretary-general), 聯合國(UN), 誰(who), 是 (was), 時(in). The first three terms (boldface type in Table 1) get more QF-Score than others because they are embodied as the titles of Wikipedia articles. We can get candidate passages and extract correct answers by these QF terms. The process of answer extraction will be described in details in Section 4.

3. Ranking Question-Focused Terms

We use the value of length-of-term and weight-in-Wikipedia to give weighted scores, which called QF-Score for candidates of QF terms. However, some terms without being embodied in Wikipedia articles may still play a significant role in a question, so we need a ranking model to differentiate the importance of these terms.

First, a candidate of QF term should have moderate length to offer enough information for extracting answers [4]. The value of length-of-term is regarded as an informative ratio of one term in the question. The ratio is calculated as Equation (1):

$$LenRatio(qf) = \frac{Length(qf)}{Length(Q)} \qquad (1)$$

where *qf* is the candidates of QF terms, *Q* is the question, *Length(qf)* and *Length(Q)* represent the length of *qf* and *Q*.

Second, we used the occurrence of a candidate in the titles of Wikipedia articles to compute its value of weight-in-Wikipedia. The similarity between candidate *qf* and Wikipedia title *T* is calculated as Equation (2):

$$Sim(qf,T) = \frac{Length(qf)}{Length(T)} \qquad (2)$$

where *T* represents the Wikipedia article titles which possibly contains candidate *qf*. If *qf* matches *T* exactly, the value of *Sim(qf,T)* would be 1, or it would be the value between 0 and 1.

As mentioned before, some terms without being embodied in Wikipedia articles may still play a significant role in the question. So we need to find another scoring standard to give them a weighted score. The weighted score is computed based on the inverse number of Google search results to decrease the weight of common terms. Therefore, we can compute the value of weight-in-Wikipedia as Equation (3):

$$Weight(qf) = \begin{cases} \sum_{T \in W} Sim(qf,T) & \text{if } qf \text{ is a part of any } T \\ \dfrac{1}{\log(Sr(qf))} & \text{otherwise.} \end{cases} \qquad (3)$$

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

where *W* is all titles of Wikipedia articles, and *Sr(qf)* represents the total number of Google search results when using *qf* as a query.

Finally we get the QF-Score by combining the value of *Weight(qf)* and *LenRatio(qf)*:

$$QF\text{-}Score(qf) = Weight(qf) \times LenRatio(qf) \quad (4)$$

We used QF-Score of each candidate of QF term to rank all candidates and the ranked result can be used for document retrieval (Section 4.2).

4. Implementing a QA System

4.1 Overview

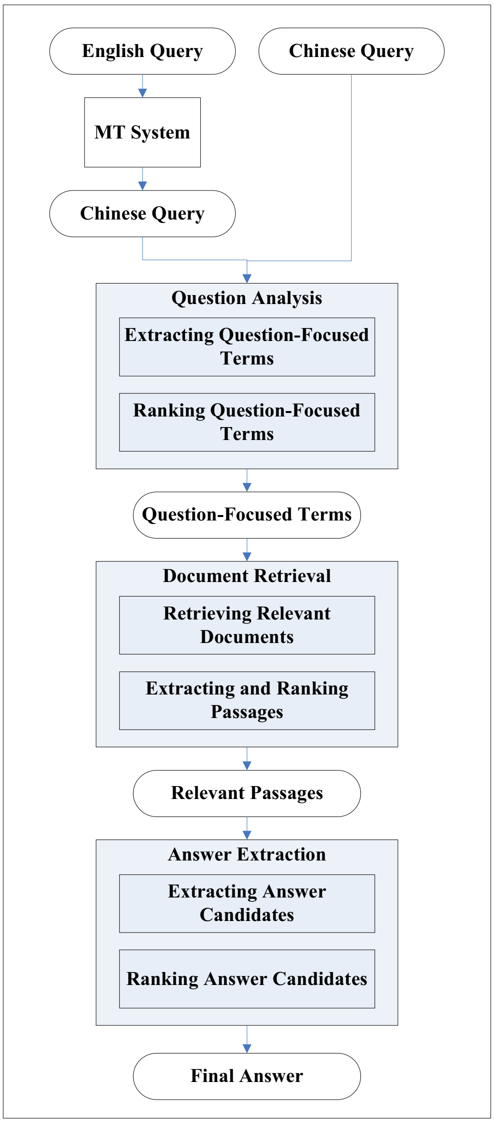
English Query
Chinese Query
MT System
Chinese Query
Question Analysis
Extracting Question-Focused Terms
Ranking Question-Focused Terms
Question-Focused Terms
Document Retrieval
Retrieving Relevant Documents
Extracting and Ranking Passages
Relevant Passages
Answer Extraction
Extracting Answer Candidates
Ranking Answer Candidates
Final Answer

Figure 1. Architecture of our QA system.

The architecture of our QA system is shown in Figure 1. It consists of three major computing modules: (1) Question Analysis, (2) Document Retrieval and (3) Answer Extraction. The only difference between C-C and E-C subtasks in our QA system is that English queries have to be translated into Chinese by a MT system. We used translated results of Google Translate, an online MT system, for our E-C subtask due to its success on terminology translation.

The module of Question Analysis has been described in details in Sections 2 and 3, so we will only describe the other two modules in the following.

4.2 Document Retrieval

Top-ranked QF terms are used to retrieve relevant documents. We assumed that the amount of retrieved relevant documents would be neither too large nor too small. Because too large amount of retrieved document would reduce the efficiency and the information from a small amount of retrieved documents may be not enough for extracting answers. Preferred amount of retrieved relevant documents may fall on the range from 10 to 1000. Therefore, we used an iterative procedure to reach the goal as Figure 2.

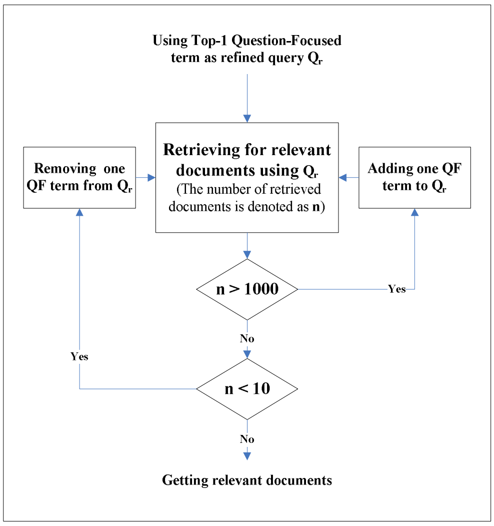
Using Top-1 Question-Focused term as refined query Q_r
Removing one QF term from Q_r
Retrieving for relevant documents using Q_r (The number of retrieved documents is denoted as n)
Adding one QF term to Q_r
n > 1000
Yes
No
Yes
n < 10
No
Getting relevant documents

Figure 2. Procedure of retrieving relevant documents.

Q_r is a refined query which contains Top-k ranked QF terms to retrieve relevant documents. QF terms are added or removed from Q_r based on the number of documents retrieved by Q_r. If the amount of retrieved relevant documents could not be adjusted to the range from 10 to 1000, the number of terms in refined query Q_r would be set to a minimum (1) or a maximum (the number of all terms in a question).

After getting relevant documents, we split them into passages using some punctuation marks. Then all passages are ranked by the summation of QF-Score of all QF terms involved in them.

$$Score\,(P) = \sum_{qf \in P} QF\text{-}Score\,(qf) \qquad (5)$$

where *P* is a passage of relevant documents and *qf* denotes a QF term.

4.3 Answer Extraction

The relevant passages are segmented into terms by CKIP tagger. We think that most answers would be tagged as nouns, including person names, location names, numbers, dates and other proper nouns. Thus only the segmented terms with being tagged as nouns could become answer candidates.

In general, most methods to rank answer candidates usually based on the scores computed using the distance between answer candidates and keywords [6, 7]. We computed the scores of answer candidates by combining the scores of passages, the scores of QF terms and the distance between answer candidates and QF terms as Equation (6):

$$Score(A) = \sum_{qf \in P} \frac{Score(P) \times QF\text{-}Score(qf)}{D(A, qf)} \qquad (6)$$

where *A* is an answer candidate and *D(A, qf)* represents the distance between *A* and a QF term *qf*.

In practice, this method may spend too much time to compute the score of each answer candidate in the passage. Therefore, we have adjusted our system to compute the score of an answer candidate whose distance with QF terms is less than 3.

Table 2. Examples of Answer Extraction.

Candidate passage #1	安南之前的多數**聯合國秘書長**想必不至於採取他的做法，問題是盧安達、科索伏、東帝汶等地的腥風血雨已經使安南成為一位深具使命感的**聯合國秘書長**。
Answer candidates	多數, 使命感
Candidate passage #2	在**聯合國**方面、**聯合國**官員說，**聯合國秘書長**安南目前正考慮是否前往伊拉克訪問，**聯合國**一個代表團則將於明天前往伊拉克觀察伊拉克當局不准**聯合國**武器檢查員進入的八個地點。
Answer candidates	安南, 方面, 官員, 當局, 武器, 檢查員

Table 2 shows examples of answer candidate extraction which follow the example in Table 1. The correct answer “安南” (Kofi Annan) which appears in passage #1 could not become an answer candidate because the distance is not less than 3. However, in passage #2, “安南” gets a highest score over all candidates in both passages #1 and #2. The reasons include: (1) its position is right after the QF terms “聯合國” (UN) and “秘書長” (secretary-general) and (2) the score of passage #2 is higher than the score of passage #1. The ranked results of answer candidates are shown in Table 3.

Table 3. Examples of Answer Ranking.

Rank	Candidate	Score
1	**安南**	**0.0405600050008003**
2	方面	0.0308903561014772
3	多數	0.0249926946759371
4	武器	0.0247122848811817
5	官員	0.0247122848811817
6	檢查員	0.00617807122029543
7	當局	0.00617807122029543
8	使命感	0.00463575482399044

5. Experimental Results

Our experimental results showed that the accuracy of Top-1 was only about 0.153 at the C-C subtask and 0.067 at E-C subtask. After we observed and analyzed the candidates of answers, we found that the Top-10 accuracy achieved over 0.50 on which question type was classified as PERSON or LOCATION. However, our system was ineffective for questions with numeric answers. This result may be caused by the insensitivity of dates and numbers both in our QA system and the titles of Wikipedia articles.

5.1 C-C Subtask

Table 4 shows the performance of our system for each question type at C-C subtask. PERSON and LOCATION are the question types with better Top-1 accuracy while no correct answer in Top-1 candidate is found with these question types: ARTIFACT, MONEY, NUMEX, PECENT and TIME.

Table 4. Evaluation report of our formal run at C-C subtask.

Results of Right + Unsupported Answers			
QType	#Q	#Correct Top-1	Ratio
ARTIFACT	7	0	0.0000
DATE	39	2	0.0513
LOCATION	**16**	**6**	**0.3750**
MONEY	8	0	0.0000
NUMEX	11	0	0.0000
ORGANIZATION	**16**	**3**	**0.1875**
PERCENT	4	0	0.0000
PERSON	**47**	**12**	**0.2553**
TIME	2	0	0.0000
Total	150	23	0.1533

We found that most numeric answer candidates are ranked with lower scores. The reasons are as follows:

(1) No employing question type classification.

(2) No numeric terms embodied in the titles of Wikipedia articles except some particular years or dates.

However, the Top-1 accuracy for non-numeric answers is lower than our expectations. In our analysis, we will focus on three question types, PERSON, LOCATION and ORGANIZATION.

Figure 3 shows the accuracy of Top-N answer candidates for PERSON, LOCATION and ORGANIZATION. Top-10 accuracy achieved 0.57, 0.5 and 0.44 respectively. Moreover, Top-50 accuracy achieved 0.74, 0.63 and 0.56 respectively. The results represent that our QA system could extract candidates from retrieved passages correctly, but the ranking model should be improved. Maybe more parameters of characteristics of answer candidates could be added into the model.

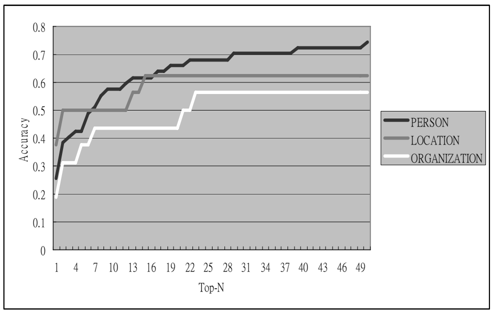

Figure 3. Top-N answer candidates accuracy for three question types.

Table 5. Examples of segmentation errors from CKIP tagger.

QID	**CLQA2-ZH-T3016-00**
Question	一九九九年時南韓總統是誰？(Who was the president of Korea in 1999?)
Correct answer	**金大中 (Kim Dae Jung)**
Segmented passage	**南韓**, **總統**, **金大(N)**, **中(N)**, 與, 日本, 首相, 小, 淵惠, 三廿日, 在, 南韓, 青瓦臺, 舉行, 會談
Top-1 answer	**金大 (Kim Dae)**
QID	**CLQA2-ZH-T3138-00**
Question	誰是泰國總理？ (Who is the prime minister of Thailand?)
Correct answer	**乃川 (Chuan Leekpai)**
Segmented passage	**泰國**, **總理**, **乃(V)**, **川(N)**, 表示, ，, 經濟, 問題, 可能, 使, 泰國, 無法, 符合, 上, 項, 方案, 附帶, 的, 一, 項, 重要, 條件,
Top-1 Answer	**川 (Chuan)**

Another reason of low accuracy is incorrect segmentation from CKIP tagger. Some proper nouns may be segmented to two or more terms. Table 5 shows examples of segmentation errors.

The first example in Table 5 shows that the name of the president of Korea "金大中" (Kim Dae Jung) has been divided into two nouns as "金大" (Kim Dae) and "中" (Jung). The other example shows that "乃川" has been divided into one verb "乃" (is) and one noun "川" (Chuan). The Top-1 answer is "川" (Chuan) because only nouns could become answer candidates.

5.2 E-C Subtask

At E-C subtask, we used Google Translate as our translation method. All other processes of our E-C subtask are the same as C-C subtask. The performance of each question type at E-C subtask is shown in Table 6. The accuracy is nearly one half of that of our C-C subtask.

Table 6. Evaluation report of our formal run at E-C subtask

Results of Right + Unsupported Answers			
QType	#Q	#Correct Top-1	Ratio
ARTIFACT	7	0	0.0000
DATE	39	1	0.0256
LOCATION	**16**	**2**	**0.1250**
MONEY	8	0	0.0000
NUMEX	11	0	0.0000
ORGANIZATION	16	1	0.0625
PERCENT	4	0	0.0000
PERSON	**47**	**6**	**0.1277**
TIME	2	0	0.0000
Total	150	10	0.0667

Table 7. Examples of Google Translate results.

QID	**CLQA2-ZH-T3146-00**
Question	Who is the **director** of "**Holy Smoke**"?
Translated result	擔任**教導主任**"**羅馬濃煙**"?
Wrong translation	Director → 教導主任 (the head-teacher) Holy Smoke → 羅馬濃煙(Rome smoke)
Correct translation	Director → 導演 Holy Smoke → 聖煙烈情
Top-1 answer (wrong)	國小 (elementary school)
QID	**CLQA2-ZH-T3147-00**
Question	Who was the **Israeli Prime Minister** in 1998?
Translated result	曾任**以色列總理**,1998?
Correct translation	Israeli Prime Minister → 以色列總理
Top-1 answer (correct)	內唐亞胡 (Benjamin Netanyahu)

Table 7 shows two examples of results of question translation using Google Translate. One gets incorrect translation but the other is translated correctly. However, the correctness of translation results plays an important role in a CLQA system.

6. Conclusion

In this paper, we have utilized a new resource, Wikipedia, for extracting and ranking QF terms to complete a CLQA system. The experimental results shows that the resource is effective, but the model of answer ranking should be improved. We also conducted a CLQA experiment with an online MT system, Google Translate, and get 0.067 Top-1 accuracy which is nearly one half of that of our C-C subtask. In the future, we will develop our own segmentation and translation tools to improve our CLQA system.

7. Reference

[1] Shu-Jung Lin, Min-Shiang Shia, Kao-Hung Lin, Jiun-Hung Lin, Scott Yu, Wen-Hsiang Lu. Improving Answer Ranking Using Cohesion between Answer and Keywords. *Proceedings of the NTCIR5 Workshop*, 2005.

[2] Yutaka Sasaki, Question Answering as Question-Biased Term Extraction: A New Approach toward Multilingual QA. *Proc. of ACL-2005*, pp.215-222, 2005.

[3] Yutaka Sasaki, Baseline Systems for NTCIR-5 CLQA1: An Experimentally Extended QBTE Approach. *Proceedings of the NTCIR5 Workshop*, 2005.

[4] Cheng-Wei Lee, Cheng-Wei Shih, Min-Yuh Day, Tzong-Han Tsai, Tian-Jian Jiang, Chia-Wei Wu, Cheng-Lung Sung, Yu-Ren Chen, Shih-Hung Wu, Wen-Lian Hsu. ASQA: Academia Sinica Question Answering System for NTCIR-5 CLQA. *Proceedings of the NTCIR5 Workshop*, 2005.

[5] K. Tanaka Ishii, H. Nakagawa. A Multilingual Usage Consultation Tool Based on Internet Searching -More than a Search Engine, Less than QA-. *WWW Conference*, pp.363-371, 2005.

[6] Naoya Hidaka, Fumito Masui, Keiko Tosaki. MAIQA: Mie Univ. Participated System at NTCIR4 QAC2. *Proceedings of the NTCIR4 Workshop*, 2004.

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