Finding Answers to Definition Questions Using Web Knowledge Bases *

Han Ren[a], Donghong Ji[a], Jing Wan[b], and Chong Teng[a]

[a]School of Computer Science, Wuhan University
129 Luoyu Road, Wuhan 430079, China
cslotus@mail.whu.edu.cn, donghong_ji@yahoo.com, tengchong@whu.edu.cn
[b]School of Chinese Language & Literature, Wuhan University
129 Luoyu Road, Wuhan 430079, China
jennifer_wanj@yahoo.com.cn

Abstract. Current researches on Question Answering concern more complex questions than factoid ones. Since definition questions are investigated by many researches, how to acquire accurate answers still becomes a core problem for definition QA. Although some systems use web knowledge bases to improve answer acquisition, we propose an approach that leverage them in an effective way. After summarizing definitions from web knowledge bases and merge them to a definition set, a two-stage retrieval model based on Probabilistic Latent Semantic Analysis is produced to seek documents and sentences in which the topic is similar to those in definition set. Then, an answer ranking model is employed to select both statistically and semantically similar sentences between sentences retrieved and sentences in definition set. Finally, sentences are ranked as answer candidates according to their scores. Experiments indicate following conclusions: 1) specific summarization technologies improves definition QA systems to a better performance; 2) topic based models can be more helpful than centroid-based models for definition QA systems in solving synonym and data sparse problems; 3) shallow semantic analysis is effective to find discriminative characteristics of definitions automatically.

Keywords: web knowledge bases, question answering, definition questions, probabilistic latent semantic analysis.

1 Introduction

Current researches on Question Answering (QA) mainly concern more complex questions than factoid ones. In TREC2007, ciQA track (Dang *et al*., 2007) focuses on 'Relationship' questions, which is defined as the ability of one entity to influence another, including both the means to influence and the motivation for doing so. In NTCIR-7, complex question (Mitamura *et al*., 2008) are taxonomically defined as four types (Event, Definition, Biography and Relationship). Sentences that contain correct responses are extracted as answer candidates. In most cases, the correct answer for a complex question is composed of multiple sentences.

Complex questions refer to complex relations between semantic concepts or synthesizing processes of deep knowledge; they implicate rich information need. Take a question-answer pair in Table 1 as an example. The question is a definition one; the information need is

* This research is supported by Natural Science Foundation of China(Grant Nos. 60773011, 90820005), Wuhan University 985 Project: Language Technology and Contemporary Social Development(Grant No.985yk004), Independent Research Foundation of Wuhan University: A research on Chinese question answering based on text entailment and Hubei Province Social Science Fund: Constructing answer patterns for Chinese question answering system.

supposed to be the consequence between the greenhouse gas and the greenhouse effect, and names of gases that cause the greenhouse effect. Therefore, Answer 1 and Answer 2 are both correct answers for the question.

Table 1: A definitional question-answer pair.

Question	*What is the greenhouse gas*?
Answer 1	*Greenhouse gas is a kind of gas that causes greenhouse effect*.
Answer 2	*Greenhouse gases include carbon dioxide, methane, nitrous oxide, etc*..

Questions like "*What is the greenhouse gas*?" or "*Who is Kofi A. Annan*?" are assigned to definition questions, and this type of questions has become especially interesting due to their high frequency in real user logs (Figueroa, 2009). Similar to complex questions of other types, the answer to a definition question is a combination of complex semantic relations and should be accurate as while as non-redundant. Thus how to acquire precise information need becomes a core problem for definition QA.

In order to acquire accurate answers (or nuggets), many systems aim at patterns of definition sentences. Such approaches leverage the syntactic styles of definition sentences and convert them to patterns for answer sentence retrieval. We (Ren *et al*., 2008) built a pattern list for Chinese complex questions by manual work. Wu *et al*. (2008) extracted definitional patterns from the Wikipedia data using regular expressions. Cui *et al*. (2004a) employed soft patterns (also known as probabilistic lexico-syntactic patterns), which were produced by unsupervised learning. These approaches are supported either by manual work or by annotated corpus more or less.

Some approaches seek another access to solve the problem. Harabagiu *et al.* (2006) decomposed complex questions to factoid ones using lexico-semantic resources. Answers to the factoid questions are fused as the answer to the original question. Such methods convert information need acquisition to decomposition problem, by which abstract information need can be changed to concrete concepts. But general lexico-semantic resources may result in lower performance against specific resources like dictionaries.

For a better performance, most systems employ specific knowledge bases such as Wikipedia or Encarta; and the essential reason is that the specific knowledge involves almost entire information need for a definition question. Hickl *et al*. (2007) searched the original complex questions into Wikipedia and calculate similarity between sentences in Wikipedia and the corpus. Zhang *et al*. (2005) utilized multiple web knowledge bases to improve EAT acquisition. Cui *et al*. (2004b) also indicated that Specific Web knowledge gleaned from definitional Web sites greatly improves the performance of definitional QA. However, most of these systems may achieves low performances, since complex questions imply relations (such as semantic similarity) between terms whereas these systems do not take them into account and just statistically retrieve and rank documents or sentences by centroid-based (or bag-of-words) method.

In this paper, we propose an approach based on web knowledge bases. Our work differs from those mentioned above is that we employ web knowledge bases with an effective way. Since we are not meant to utilize any other method by manual work or annotated corpus, web knowledge bases are the exclusive way for our approach to acquire information need of questions. First, we summarize definitions from web knowledge bases and merge them to a definition set; then a two-stage retrieval model based on Probabilistic Latent Semantic Analysis (PLSA) is produced to seek documents and sentences in which the topic is similar to those in definition set; after that, an answer ranking model is employed to extract sentences which not

only statistically but also semantically similar with any sentence in definition set. Finally, sentences are ranked as answer candidates according to their scores. Experiments in NTCIR-7 data set show that our approach leverages web knowledge bases effectively and achieves better performance than the baseline system in NTCIR-7.

The rest of the paper is organized as follows. In Section 2, we show the processing mechanism of our system. In Section 3, we give methods of definition acquisition and answer ranking in detail. In Section 4, we discuss the experimental results. Finally, the conclusion and future work are given in Section 5.

2 System Overview

Our system utilizes a general QA framework which mainly contains three models: text annotation, document retrieval and answer ranking. To obtain the definition set, a summarization model is also employed. The system shown in Figure 1 carries out the following steps which are briefly described as follows.

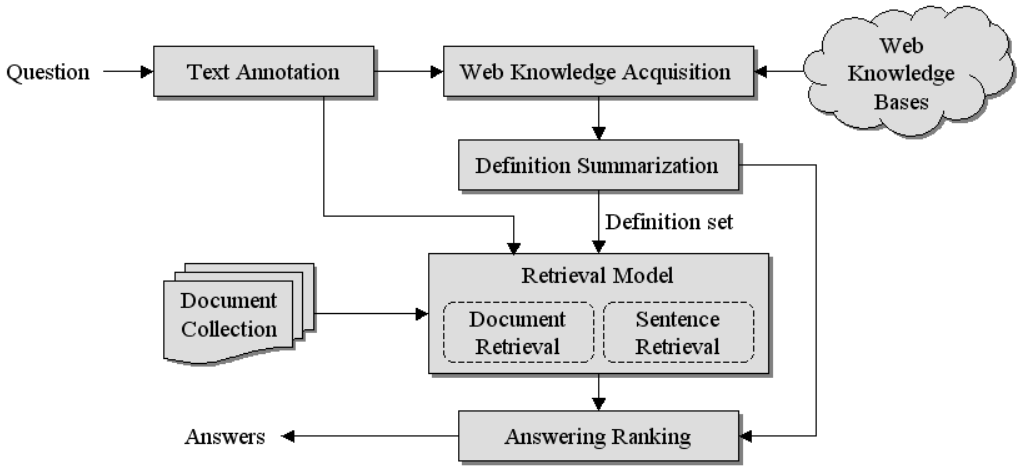

Figure 1: System processing mechanism.

In text annotation model, words in questions and definitions summarized from web knowledge bases are segmented since the data set we utilize in our experiments is a simplified Chinese version. After eliminating stop words, we recognize named entities from sentences of questions and definitions. Finally, named entities and other terms are put into web knowledge bases for retrieval.

We employ two summarization models for our system. The first one is an effective model that summarizes definitions from Wikipedia and the other one is a traditional single document summarization model. We first summarize definitions from Wikipedia and get a basic definition set. Then we utilize the traditional model to deal with other web knowledge bases and get a extent definition set. After that we combine the extent set to the basic set without sentence redundancy.

For document retrieval, we utilize a statistical IR model based on VSM to retrieve documents in a wide range. Then we produce a topic model based on PLSA to compare topical similarity between definition set and sentences in documents.

In answer ranking phase, sentences in relevant documents are weighted with those in definition set. The weighting method is based on statistical and shallow semantic similarity. Sentences are ranked and those with high scores are picked out as answer candidates.

In this paper, we mainly focus on how to utilize PLSA model to improve IR performance, and how to acquire precise answers by web knowledge bases. Thus details of retrieval model and answering ranking will be introduced in the following section.

3 Definition Question Answering Based on Web Knowledge Bases

In the experiments we utilize three web knowledge bases that are Wikipedia[1], Baidupedia[2] and Hudong[3]. All of them provide simplified Chinese version, so that we can easily use them for the experiments.

3.1 Definition Summarization

Although web knowledge bases are cleaned up and generalized by manual work, they still permeate insignificant information that may decrease the performance of document retrieval or answering ranking. To solve this problem, some systems make use of summarizing methods to acquire important portions from definitions. But when definitions are not rich, the snippets summarizing from them are also sparse. Since some web knowledges such as Wikipedia provide links to combine enormous concepts, documents linked by these links allows us to obtain more reliable texts. Ye *et al*. (2009) proposed a novel approach that can produces summaries with various length. By building an extended document concept lattice model, concepts and non-textual features such as wiki article, infobox and outline are combined. Experiments showed that system performance outperformed not only traditional summarizing methods but also some soft-pattern approaches. In this paper, we utilize this approach to perform definitions in Wikipedia. Definitions summarized are put into basic definition set.

Although most definitions can be found in Wikipedia, we still utilize other two web knowledges as a supplement. Following an unsupervised summarization approach proposed by Ji (2006) to rank sentences, text summarized are put into extend definition set. Then we add sentences in the extent definition set to the basic set without sentence redundancy.

3.2 PLSA based Document Retrieval

Traditionally, retrieving a definition question may encounter more difficult than other complex questions because information derived from a definition question is quite insufficient for retrieval. For instance, the question in Table 1 has only one concept *greenhouse gas*, which is much difficult for IR models to seek relevant documents. To solve it, most QA systems employ various external resources such as WordNet or search engines to expand queries. But the performance of these systems can not reach those of which utilize specific resources like Wikipedia. For this reason, we consider using definitions summarized from web knowledge bases to expand queries formed by concepts in definition questions.

As to IR models, many of them consider documents or sentences as a bag of words, which can provide a good performance at document level retrieval. But for QA systems, they encounter two problems that can not increase the performance effectively: (1) synonym and polysemy of concepts; and (2) data sparse in small text retrieval. Different from methods above, topic models such as PLSA build a semantic (or topic) layer to combine words and documents. By casting them into a semantic layer, synonyms and data sparse problem can be overcome in a certain extent. Following is the description of our PLSA model for sentence retrieval.

Given a sentence set S, a term set W and a topic set Z, the conditional probability of sentence-term $P(s, w)$ can be described as follows:

$$P(s,w)=\sum_{z\in Z}P(z)P(s\mid z)P(w\mid z) \tag{1}$$

[1] http://zh.wikipedia.org/
[2] http://baike.baidu.com/
[3] http://www.hudong.com/

$P(w \mid z)$ represents the conditional probability of words in latent semantic layers (or topics), $P(z \mid s)$ represents the conditional probability of topics in sentences. Here the count for topic set Z is between 20 and 100. Then the model fits with the EM algorithm and export the optimal $P(Z)$, $P(W \mid Z)$ and $P(Z \mid S)$. When a new query is coming, it is projected to the topic space Z. The similarity of the query and each sentence can be acquired by computing the similarity of the probabilistic distribution between them in the topic space.

In this paper, we adopt two-stage retrieval strategy, that is, first a VSM based model is utilized to retrieve relevant documents in a wide range. All of the documents can be treated as a sentence set. Then a PLSA based retrieval model is employed by retrieval model. For each sentence in definition set, it is treated as a query and submitted to the model to seek relevant sentences in sentence set. Sentence that its value achieves a threshold are picked out as the sentence candidates.

3.3 Answering Ranking

Our method of answer ranking considers syntactic/semantic and statistical information of sentences. On the one hand, although sentences retrieved have the latent semantic similarity with definition sentences from web knowledge bases, they are probably more similar with definition sentences if they have the syntactic or actual semantic similarity. Moreover, we can easily extract nuggets by using the same semantic constituents. On the other hand, to balance the impact of syntactic/semantic judgment, we utilize a statistical similarity which treats sentences as a bag of words. The motivation is, a sentence can be a potential answer candidate if most of the words in it are also appear in definitions derived from web knowledge bases.

For semantic similarity of definitions, we consider the similarity of the core semantic roles, which primarily profile the features of definitions. We choose the verb based labeling architecture derived from PropBank in which 'predicate', 'subject' and 'object' are the core roles for a sentence. For each sentence in definition set we only label PRED, A0 and A1 and combine them to a triple. For example, the triple of Answer 1 in Table 1 is:

$$\{is \mid \text{PRED},\ Greenhouse\ gas \mid \text{A0},\ gas \mid \text{A1}\}$$

For labeling of semantic roles, we utilize a method which we implemented in CoNLL shared task (Ren *et al*., 2009) to extract triples. They handle syntactic dependency parsing with a transition-based approach and utilize MaltParser [4] as the base model. We also utilize a Maximum Entropy model to identify predicate senses and classify arguments.

Although definitions summarized from web knowledge bases are more precious, there still have some sentences that are not definitions. Since a definition sentence quite probably has the target of the definition, triples can only be acquired by these sentences. Also, co-reference resolution is made use of to help find actual definition sentences in web knowledge bases. Finally, sentences that do not contain the target of a definition are removed from the definition set.

The wegithing formula of our method for answering ranking is as follows:

$$R(s_i) = \alpha \bullet max\{W(s_i, p_j)\} + (1-\alpha) \bullet S(s_i, k) \quad (2)$$

W and S denotes the semantic and statistical similarity respectively. s_i represents sentence i, p_j represents triple j, and k means definition set. In the formula, α is an adjusting parameter.

W is a binary function that is described below:

[4] http://w3.msi.vxu.se/~jha/maltparser/

$$W(s_i, p_j) = \begin{cases} 1, & \text{one or more roles are same except target} \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

W means that if at least two roles in s_i and p_j are same except the target (eg. A0 and the predicate), the value is 1; otherwise the value is 0.

For statistical similarity, we simply use *cosine* similarity to deal with sentence candidates and web knowledge bases and s_i and k_j are treated as a bag of words.

4 Experiments and Analysis

4.1 Experiment Setting

We utilize Chinese data sets of NTCIR-7 CCLQA as our data set. The data set includes 545,162 documents which come from Xinhua Newspaper and Lianhe Zaobao. The evaluation data contains 20 definition questions and biography ones respectively. We treat these 40 questions as definition ones.

For each question we retrieve the concept in it using three web knowledge bases and merged to one definition text. Thus we collect 40 texts that include definitions for questions. In the experiment we tune summarization models to export 50 sentences for each question.

For the first retrieval by VSM based IR model we select top 1000 documents and use them for latent semantic retrieval. Empirically, we set topic number in PLSA model as 50.

In answer ranking, the adjusting parameter α is initially set as 0.5, which means that we treat semantic and statistical similarity equally.

4.2 Evaluation Metrics

In the experiment, we utilize the method described in NTCIR-7 for the evaluation. The method adopts F-score as the evaluation index where β is a parameter signifying the relative importance of precision and recall. In order to get comparable results with baseline system, β value is fixed to 3.

$$F_{\beta=3} = \frac{(\beta^2 + 1) \times precision \times recall}{\beta^2 \times precision + recall} \quad (4)$$

Parameters that appear at Formula (4) are shown in Figure 2.

r	sum of weights over matched nuggets
R	sum of weights over all nuggets
a	# of nuggets matched in system response
L	total character-length of system response
C	character allowance per match
allowance	$a \times C$
recall	r / R
precision	$\begin{cases} 1 & \text{if L < allowance} \\ \frac{allowance}{L} & \text{otherwise} \end{cases}$

Figure 2: Parameters in evaluation method.

4.3 Experimental Results and Analysis

We select NTCIR-7 organizer's system as the baseline system in our experiments. The system makes use of the same architecture with JAVELIN (Nyberg *et al.*, 2003), whereas does not take answer types into account and simply extracts noun phrases from questions as key terms. For answer extraction and ranking, the system selects sentences that contain key terms from high ranked documents. Table 2 shows the result of the experiment.

Table 2: Overall Performance (%).

	Definition	Biography	Average
this paper	**24.31**	**20.67**	**22.49**
baseline	13.60	12.48	13.04

The results indicate that our approach based on web knowledge bases can improve the overall performance of QA system. We can also see that, result of definition is more better than that of biography. It is mainly because named entities in biography texts are more than those in definition ones. For example, a temporal named entity almost appears at every biographic text whereas the entity does not represent a rich semantic concept; the PLSA retrieval model can not make a relationship with the entity and a topic in layer *Z*. Thus the sentence involving this may not have a tight relationship with definitions from web knowledge bases although the sentence could be an answer.

We also investigate performance of each part in our system. For summarizing definitions, we utilize two methods proposed in Section 3.1, namely a definition summarization(ds) and an unsupervised summarization(us). For sentence retrieval, two models, which are based on VSM and PLSA mentioned in Section 3.2, are also involved in the experiments. For answer ranking, as a comparison, we invoke the approach of the baseline system that rank sentences by key terms to replace our method proposed in Section 3.3(sem). Table 3 shows the result of the experiments.

Table 3: Comparsion of system performances based on different parts (%).

	Definition	Biography	Average
VSM+key term	11.95	9.09	10.52
PLSA+key term	13.68	10.71	12.20
PLSA+sem	19.49	14.73	17.11
us+PLSA+sem	21.03	16.56	18.80
ds+PLSA+sem	**22.67**	**18.98**	**20.83**

From Table 2 we can see that when adopting the definition summarization method, the performance increases 1.64% of definition questions and 2.42% of biography ones in contrast with the method of unsupervised summarization. When using VSM model to replace PLSA, the performances decrease 1.73% and 1.62% respectively. In addition, the average performance noticeably decreases 4.91% by using key term method to replace the method of computing semantic plus statistical similarity. Data show the facts that: (1) effective summarizing definitions from Wikipedia improves ordinary QA systems to a better performance; (2) a large number of synonyms and polysemys as well as data sparse phenomenon exist in texts, thus topic based models can be more fit than centroid-based models for complex QA systems; (3) a sentence that just contains key terms is not considered as a definition unless it has pattern or semantic features that definition bears.

In order to investigate the effect of semantic or statistical similarity, we set another experiment by adjusting α from 0 to 1. Figure 3 shows the results. We can see that when α = 0.6 the system achieves the best performance. We also realize that when α changes from 0.6 to 0, the performance decrease sharply than the performance when α changes from 0.6 to 1. It indicates that the semantic similarity is a little more important than the statistical one. More specifically, the impact of methods that consider terms and their relations is more notable than that of those methods that just take texts as a bag of words for answer identification and ranking of definition questions.

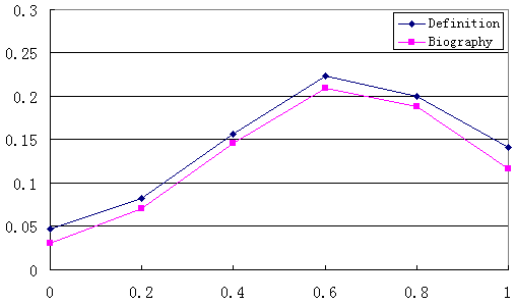

Figure 3: Adjusting α parameter.

5 Conclusion

Current researches on Question Answering mainly concern more complex questions than factoid ones. In this paper, we propose an approach to leverage web knowledge bases effectively. After summarizing definitions from web knowledge bases and merging them, a two-stage retrieval model based on Probabilistic Latent Semantic Analysis is employed seek documents and sentences in which the topic is similar to that in definition set. Finally, an answer ranking model is utilized to rank both statistically and semantically similar sentences between sentences retrieved and sentences in definition set. Experiment shows that our system yields a better performance than the official one of NTCIR-7. In the future, we aim at improvement of more effective topic models, which could achieve a better performance in dealing with complex question answering.

References

Cui, Hang, Min-Yen Kan and Tat-Seng Chua. 2004a. Unsupervised Learning of Soft Patterns for Generating Definitions from Online News. In *Proceedings of Thirteenth World Wide Web Conference*. pp.90-99.

Cui, Hang, Min-Yen Kan, Tat-Seng Chua and Jing Xiao. 2004b. A Comparative Study on Sentence Retrieval for Definitional Question Answering. In *Proceedings of SIGIR 2004 Workshop on Information Retrieval for Question Answering*.

Dang, Hoa Trang, Diane Kelly and Jimmy Lin. 2007. Overview of the TREC 2007 Question Answering Track. In *Proceedings of the Sixteenth Text Retrieval Conference*. NIST, Gathersburg.

Figueroa, Alejandro. 2009. Finding Answers to Definition Questions across the Spanish Web. In *Proceedings of the Eighteenth World Wide Web Conference*. Madrid, Spain.

Harabagiu, Sanda, Finley Lacatusu and Andrew Hickl. 2006. Answering Complex Questions with Random Walk Models. In *Proceedings of SIGIR'06*. Seattle, Washington.

Hickl, Andrew, Kirk Roberts, Bryan Rink, Jeremy Bensley, Tobias Jungen, Ying Shi and John Williams. 2007. Question Answering with LCC's CHAUCER-2 at TREC 2007. In *Proceedings of the Sixteenth Text Retrieval Conference*. NIST, Gathersburg.

Ji, Paul. 2006. Multi-document Summarization Based on Unsupervised Clustering. In *Proceedings of the Third Asia Information Retrieval Symposium*. Singapore.

Mitamura, T., E. Nyberg, H. Shima, T. Kato, T. Mori, C.Y. Lin, R.H. Song, C.J. Lin, T. Sakai, F. Gey, D.H. Ji and N. Kando. 2008. Overview of the NTCIR-7 ACLIA: Advanced Cross-Lingual Information Access. *NTCIR-7*, Tokyo.

Nyberg, E., T. Mitamura, J. Callan, J. Carbonell, R. Frederking, K. Collins-Thompson, L. Hiyakumoto, Y. Huang, C. Huttenhower, S. Judy, J. Ko, A. Kupść, L.V. Lita, V. Pedro, D. Svoboda and B. Van Durme. 2003. The JAVELIN Question-Answering System at TREC 2003: A Multi-Strategh Approach with Dynamic Planning. In *Proceedings of the Twelfth Text Retrieval Conference*.

Ren, Han, Donghong Ji, Yanxiang He, Chong Teng and Jing Wan. 2008. Multi-Strategy Question Answering System for NTCIR-7 C-C Task. In *Proceedings of the Seventh NTCIR Workshop Meeting*. Tokyo, Japan.

Ren, Han, Donghong Ji, Jing Wan and Mingyao Zhang. 2009. Parsing Syntactic and Semantic Dependencies for Multiple Languages with A Pipeline Approach. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, Boulder, Colorado.

Wu, Youzheng, Wenliang Chen and Hideki Kashioka. 2008. NiCT/ATR in NTCIR-7 CCLQATrack: Answering Complex Cross-lingual Questions. In *Proceedings of the seventh NTCIR Workshop Meeting*. Tokyo, Japan.

Ye, Shiren, Tat-Seng Chua and Jie Lu. 2009. Summarizing Definition from Wikipedia. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics*. Singapore.

Zhang, Zhushuo, Yaqian Zhou, Xuanjing Huang and Lide Wu. 2005. Answering Definition Questions Using Web Knowledge Bases. In *Proceedings of IJCNLP2005*. Korea.