Using Multiple Combined Ranker for Answering Definitional Questions

Junkuo Cao, Lide Wu, Xuanjing Huang, Yaqian Zhou, and Fei Liu

Fudan University, Shang Hai, China
{jkcao,ldwu,xjhuang,zhouyaqian, liufei}@fudan.edu.cn

Abstract. This paper presents a Multiple Combined Ranker (MCR) approach for answering definitional questions. Generally, our MCR approach first extracts question target-related knowledge as much as possible, then using this knowledge to pick up appropriate question answers. The knowledge includes both online definitions and related terms (RT). In our system, extraction of related terms is different from traditional methods which are largely based on calculating the co-occurred frequency of target words. We adopted the significance of sentences and documents, from which RT were extracted. The MCR approach shows state-in-art performance in handling with increasingly complex definitional questions.

Keywords: Definitional question answering, Multiple Combined Ranker, Related Terms Extraction.

1 Introduction

The objective of question answering task is to focus research on systems that return merely answers, rather than documents containing answers. Related work concerning to definitional question answering are mostly concentrated on Patterns Extraction, Ccentroid-based ranking, as well as utilizing Web knowledge as external source. Patterns Extraction has been extensively adopted in information retrieval tasks. These patterns are often expressed and matched against as regular expressions. Sudo et al. [1] employed TF*IDF to get a set of relevant sentences and built patterns from them. Other approaches employed to extract definitional sentences include various pattern matching methods, in which hand-crafted or machine learned rules are generated to find nuggets[2][3][4]. Moreover, some definitional question answering systems adopt a centroid-based ranking method to identify and select definition sentences [2][5]. For each question target, a series of centroid words were identified and grouped into a centroid vector, which was utilized to rank input sentences using cosine similarity.

Our multiple combined ranker (MCR) approach for answering definitional questions differs from the above in that we perform sentence selection process in a novel and effective way. Instead of using Centroid-based or Pattern-based method, we adopt different rankers, which respectively measures candidate sentences' importance based on AQUAINT corpus, question target expansion, as well as Web knowledge collections. These three rankers act as mutual supplements.

H. Li et al. (Eds.): AIRS 2008, LNCS 4993, pp. 448 – 453, 2008.

2 Multiple Combined Ranker

2.1 Basic Ranker

We use Basic Ranker as the first part of definitional question answering process, it consists of two components: the searching procedure and the refining procedure. A search engine (Lucene) retrieves documents in respond to the target query, and ranks them using some algorithm. We make the assumption that the search engine already produced a good result. Consequently, the sentences in these documents are supposed to have a tight relationship with the question target. The refining procedure considers other possible factors that might make a candidate sentence become an appropriate answer. The assumption is that the sentences containing words, phrases and Name Entities that co-occur frequently with the target are largely possible to be important ones for answering the question. Also, a single sentence could appear in several documents, the total number of these documents is supposed to be in direct relationship to the significance of this sentence concerning the target. Based on the above two procedures, the Basic Ranker scores the candidate sentences so that the relevant sentences receive higher scores.

2.2 Web Ranker

Until recently, Web knowledge bases (Web KBs) are increasingly recognized as a promising way to provide online knowledge, thus we adopt Web KBs as an alternative way for knowledge acquisition and build another ranker called Web Ranker. During this procedure, we calculate the similarity scores between candidate sentence and definitions from different knowledge bases respectively, and merge these scores to rank the candidate sentences. For each target, its candidate sentences are ranked using definitions from Web KBs. Firstly we construct a words vector space, which is based on TF*IDF, for all candidate sentences and Web definitions. Each of them is projected into this vector space. Secondly, the similarity of a particular candidate sentence and Web definition are computed based on the cosine of the two vectors.

Our definitional question answering systems got promising results by employing several external Web knowledge bases during TREC 2003 and TREC 2004. However, this may be due to the intrinsic simplicity of question targets more or less. But unfortunately, definitional questions have shown an increasingly complex characteristic, by way of adopting more complex question types. As a result, we could only find out online definitions for about 65% of the total question targets in TREC2005. Although we still employ Web ranker as one of our strategies to rank the candidate sentences, it is not reliable as before, more details of this approach could be referred to in [6].

2.3 Related Terms Ranker

We construct the Related Terms (RT) Ranker based on the extension of the question targets, for the purpose of obtaining more reliable and target-related information nuggets. At the heart of RT Ranker is the process of identifying and selecting words, phrases, and Name Entities, which are in tight relationship with the question targets. These terms were acquired at the end of preliminary processes like word segmentation and stemming. Also, a Relation Degree is defined to weigh the relationship between extracted terms and the question target In previous work, expansion of terms were adopted in automatic query expansion, as well as open-domain question answering [7][8]. Our approach differ from the above in that 1) Making full use of NE extraction technology which is quite helpful in identifying Related Terms. 2) Taking into account of not only the Relation Degree of the terms, but also their weights, which are related with the Basic Ranker score of the sentence that they belong to. The set of related words, phrases, and Name Entities (naturally named Related Terms) is denoted by $T=\{t_1, t_2, ..., t_n\}$. The process of Relation Degree computing is defined as below:

$$r(t_i) = \sum_j E(t_i, S_j) \times initscore\,(S_j)$$

$$\text{Where } E(t_i,S_j) = \begin{cases} 1 & t_i \in S_j \\ 0 & t_i \notin S_j \end{cases}$$

Where *initscore*(S_j) stands for the Basic Ranker score of sentence S_j. After that, $r(t_i)$ is normalized and ranked, top terms were selected as RT for further processing.

Consequently, we rank the candidate sentences based on Relation Degree of RT. Let n_w, n_p, n_e respectively represent the number of words, phrases and Name Entities that a particular sentence S' contains, and $r(w_i)$, $r(p_i)$, $r(e_i)$ denotes the Relation Degree of them, *RT_score(S')* is introduced to denote the score of this sentence according to the Related Terms Ranker, which is defined as follows:

$$RT\_score(S') = \alpha \sum_{r_w} r(w_i) \Big/ n_w + \beta \sum_{r_p} r(p_j) \Big/ n_p + \gamma \sum_{r_e} r(e_k) \ / n_e$$

According to experiments and heuristic assumptions, Name Entity should play a relative important role in RT, thus γ received a slightly higher weight than other two parameters. In our system, they are allotted to 0.3, 0.3 and 0.4 respectively to receive the optimal result.

3 Experiments

To evaluate the effectiveness of multiple combined ranker, we utilize the data set from TREC 2006 QA track, which contained 75 series as well as answer judgments. Our system official F(β=3) scores is 0.223, ranked second in all participated systems. To further compare the effectiveness of our MCR approach, we experimented on the TREC 2005 definition question set using our evaluation system, which can keep the rank when evaluates the top 10 submitted result.

The purpose of our first experiment is to judge the effectiveness of the results of document retrieval, which is the foundation of Basic Ranker and Related Terms

Ranker. In the second experiment we evaluate effectiveness of sentence selection. The purpose of the third experiment uses the Basic Ranker as a baseline, and Multiple Combined Ranker is compared with the baseline to show its effectiveness.

3.1 Effectiveness of Document Retrieval

In this part, we utilize Lucene 2.0 as our search engine and judge the returned documents by Vital and Okay nuggets recall, respectively. We vary the number of returned documents from 1 to 200 to study the effect of document number on nuggets recall. The result is listed in Table 1.

As shown from Table 1, Vital nugget recall in all TOP200 documents can achieve up to 90.0% recall and Okay nugget recall reach 81.9%, which are especially high scores. However, this higher score is achieved at the cost of precision score since returning too many sentences for a question target inevitably adds in noise information nuggets. So we also test nuggets recall on top N (1-100) returned documents, experiment results show that the Vital nuggets recall is higher than Okay nuggets recall in TOPN documents. Because the Vital nugget is more important than Okay one, our solution of document retrieval is successful. We can also see from Table 1, R(V)/N and R(O)/N decrease with the increasing of N, which is in accordance with the Basic Ranker hypothesis.

3.2 Candidate Sentences Selection Evaluation

The returned documents always contain some sentences that were not related to the question target. Therefore, discarding the noise sentences is very important. In order to evaluate the process of candidate sentence selection, we use the same method (MCR) for definitional question answering but with different candidate sentences sets. The first set is all candidate sentences without selection. Although all candidate sentences contain 90.0% Vital nuggets and 81.9% Okay nuggets, the system's F-score is only 0.187. In contrast, the other candidate sentence set is selected by Basic Ranker and some manual constructed rules. More candidate sentences were discarded in the selection process, as shown in Table 2. The Vital nugget recall and Okay nuggets recall decreased 30.3% and 45.1% respectively. However, although both Vital recall and Okay recall decreased obviously, the system performance improved 72.0%. In the same time, we try some different candidate sentence sets in our system. These

Table 1. The performance of all candidate sentences from TOPN documents of TREC2005 definitional QA. R(V) denotes Vital Nugget Recall, and R(O) denotes Okay Nugget Recall

TOPN	R(V)	R(V)/ N	R(O)	R(O)/ N
TOP1	21.1%	0.211	11.2%	0.112
TOP5	46.7%	0.093	25.6%	0.051
TOP10	54.8%	0.055	34.0%	0.034
TOP20	61.8%	0.031	44.4%	0.022
TOP50	73.1%	0.015	60.1%	0.012
TOP100	82.4%	0.008	68.4%	0.007
TOP200	90.0%	0.005	81.9%	0.004

Table 2. The effect of candidate sentences selection in definitional question answering

Candidate Answer Sentence	Size	R(V)	R(O)	F(β=3)
All sentences without selection	56100K	90.0%	81.9%	0.186
Sentences selected by Basic Ranker	1992K	62.7%	45.0%	0.320

Table3. The comparison of using three Rankers for definitional question answering

Ranking Method	F(β=3)	R(V)	R(O)	Precision
BASIC	0.272	0.388	0.213	0.097
WEB	0.264	0.381	0.199	0.097
RT	0.211	0.297	0.167	0.083
BASIC +WEB	0.311	0.460	0.240	0.105
BASIC +RT	0.305	0.433	0.225	0.108
WEB+RT	0.280	0.412	0.193	0.097
BASIC+WEB+RT (MCR)	0.318	0.467	0.250	0.111

experiments show that the confidence of question answers is determined according to the degree of candidate sentences noise. So it is difficult but crucial to balance well the Vital/Okay nuggets information with the noise information for definitional question answering.

3.3 Effectiveness of Multiple Combined Ranker

For each candidate sentence, three scores are calculated by Basic Ranker, WEB Ranker and Related Terms (RT) Ranker respectively. These scores are then applied to extract the question answers, both respectively and synthetically. In ranking process, weights of the three scores are estimated by our automatic evaluation system. Question with different target type is allocated with different weight. The performance of these ranking procedures, briefly named as BASIC, WEB and RT, have been evaluated and are shown in Table 3. As can been seen from this table, the best single solution is BASIC. This phenomenon is largely due to the fact that, the BASIC method in choosing candidate sentences is not only an important element for answering question, but it is also the foundation of WEB Ranker and RT Ranker. The third Ranker (RT) returns the worst F-measure against other two single method though, it shows competitive performance while working together with the BASIC Ranker and WEB Ranker. We can see from Table 3 that adding Related Terms Ranker to BASIC Ranker and WEB Ranker could improve the system performance up to 12% and 6% respectively, and compared with BASIC+WEB, employing Multiple Combined Ranker (MCR) could enhance system performance by 2%. Generally, the combined solution is much better than separated ones. This could be deduced from the fact that the best solution method BASIC + WEB + RT (MCR), whose F-Measure achieved 0.318, outperformed the best single solution to a great extent (about 17% improvement).

4 Conclusion

Compared with other question answering tasks, definitional question answering has more uncertain factors. There are still many divergences even among experts while answering these questions. Therefore the key of answering these questions is to find reliable knowledge related to the target. So we propose a Multiple Combined Ranker (MCR) approach to rank candidate sentences for definitional question answering. To acquire the reliable and related information, external knowledge from online websites and the related words, phrases and entities were extracted. Using these multiple knowledge, the definitional QA system can rank the candidate answers effectively.

Acknowledgements

This research was supported by the National Natural Science Foundation of China under Grant No. 60435020 and No. 60503070

References

1. Sudo, K., Sekine, S., Grishman, R.: Automatic Pattern Acquisition for Japanese Information Extraction. In: Proc. HLT 2001, San Diego, CA (2001)
2. Cui, H., Kan, M.-Y., Chua, T.-S.: Unsupervised Learning of Soft Patterns for Generating Definitions from Online News. In: Proceedings of WWW (2004)
3. Kouylekov, M., Magnini, B., Negri, M., Tanev, H.: ITC-irst at TREC-2003: the DIOGENE QA system. In: Proceedings of the Twelfth Text REtreival Conference. NIST, GAthersburg, MD, pp. 349–357 (2003)
4. Blarr-Goldensohn, S., McKeown, K.R., Schlaikjer, A.H.: A hybrid approach for QA track definitional questions. In: Proceedings of the Twelfth Text Retrieval Conference. NIST, Gathersburg, MD, pp. 185–192 (2003)
5. Radev, D., Jing, H., Budzikowska, M.: Centroid based summarization of multiple documents. In: ANLP/NAACL 2000 Workshop on Automatic Summarization, Seattle, WA, April 2000, pp. 21–29 (2000)
6. Zhang, Z., Zhou, Y., Huang, X., Wu, L.: Answering Definition Questions Using Web Knowledge Bases. In: The Proceeding of IJCNLP, pp. 498–506 (2005)
7. Echihabi, A., Hermjakob, U., Hovy, E., Marcu, D., Melz, E., Ravichandran, D.: Multiple-Engine Question Answering in TextMap. In: The Twelfth Text REtrieval Conference (TREC 2003) Notebook (2003)
8. Kwok, C., Etzioni, O., Weld, D.S.: Scaling Question Answering to the Web. In: Proceedings of the 10th World Wide Web Conference (WWW 2001), HongKong (2001)