LETTER Special Section on Data Engineering and Information Management

Improving Question Retrieval in cQA Services Using a Dependency Parser

Kyoungman BAE[†a)], *Student Member and* Youngjoong KO[†b)], *Nonmember*

SUMMARY The translation based language model (*TRLM*) is state-of-the-art method to solve the lexical gap problem of the question retrieval in the community-based question answering (cQA). Some researchers tried to find methods for solving the lexical gap and improving the *TRLM*. In this paper, we propose a new dependency based model (*DM*) for the question retrieval. We explore how to utilize the results of a dependency parser for cQA. Dependency bigrams are extracted from the dependency parser and the language model is transformed using the dependency bigrams as bigram features. As a result, we obtain the significant improved performances when *TRLM* and *DM* approaches are effectively combined.

key words: *question retrieval, cQA service, dependency, language model*

1. Introduction

Community-based Question Answering (cQA) Services, such as Naver Knowledge Search (Naver KiN), Yahoo! Answer and Baidu Zhidao, allow users to just ask any question and get answers from other users, so they can be regarded as a kind of knowledge market. Commonly, the cQA services have to construct and maintain very large archives of previous questions and their answers [1]. Since the cQA services can directly return answers to queried questions instead of a list of relevant documents, large scale of question-answer pairs have become an important information resource [2]. If a similar question and their answers are found, then users can directly obtain answers rather than a list of potentially relevant documents in the traditional IR. The retrieval task in the cQA services is to find relevant question-answer pairs for new questions posed by a user [3].

The major challenge for the question retrieval, as for most information retrieval models, is the lexical gap between queried questions and question-answer pairs [4]. For example, “What is the best way to buy PC?” and “How to get a personal computer?” are very similar questions, but they have very few words in common. The various methods for the question retrieval have been studied, such as vector space model (*VSM*) [5], *Okapi* model [6], language model (*LM*) [7] and translation-based language model (*TRLM*) [3]. *TRLM* has yielded the better performance than the traditional methods, such as *VSM*, *Okapi* and *LM* in the question retrieval. It regarded the question retrieval task as a statistical machine translation problem by using IBM model-1 [8] to learn the word-to-word translation probabilities [1], [3], [9]. Some researchers improved the *TRLM* by considering the latent topic information or the category information [10]–[14]. However, they have a limitation from the linguistics perspective. They only focus on solving the lexical gap problem and only consider unigram features that neglect dependencies between words, although a dependency plays an important role in many linguistic phenomena.

In this paper, we explore how to utilize the results of a dependency parser to improve previous question retrieval model in the cQA. Head-dependent pairs are extracted as bigram features, called dependency bigrams, from dependency analysis results by the C&C parser[*]. Then we propose the dependency based model (*DM*) by transforming the language model (*LM*) with the dependency bigrams. Finally, we combine our proposed model with *TRLM* through a linear combination. The proposed model incorporates the requirements not only on words in a unigram model (*TRLM*), but also on dependencies between words in a bigram model (*DM*). In the experiments, our proposed combination models achieved 0.6776 and 0.6703 that are relatively 4.53% and 3.51% higher performance than *TRLM* on *MAP*. In additional, the experimental results were analyzed with respect to the abovementioned advantages of the dependency bigrams. As a result, we think that the proposed dependency based model can effectively reflect the contextual information of dependencies in the question retrieval.

This paper is organized as follows. We overview related work in Sect. 2. In Sect. 3, we describe our dependency based question retrieval model in detail. In Sect. 4, we discuss the analysis of experimental results. Finally, we draw some conclusions in Sect. 5.

2. Related Work

A lot of researchers have studied to improve question retrieval in cQA. Most of these work focus on finding semantically equivalent or close to the queried question. The unigram language model (*LM*) has been widely used for question retrieval on the cQA [7], [15].

$$P_{LM}(q\,|\,Q) = \prod_{w\in q}(1-\lambda)P_{ml}(w\,|\,Q) + \lambda P_{ml}(w\,|\,Cor) \quad (1)$$

Manuscript received June 22, 2016.
Manuscript revised October 31, 2016.
Manuscript publicized January 17, 2017.

[†]The authors are with Dong-A University, 840, Hadan 2-dong, Saha-gu, Busan, 604–714, Republic of Korea.
a) E-mail: kyoungman.bae@gmail.com
b) E-mail: youngjoong.ko@gmail.com (Corresponding author)

DOI: 10.1587/transinf.2016DAL0001

[*]http://svn.ask.it.usyd.edu.au/trac/candc/

$$P_{ml}(w\,|\,Q) = \frac{tf(w,Q)}{|Q|}, \quad P_{ml}(w\,|\,Cor) = \frac{tf(w,Cor)}{|Cor|},$$

where q is the queried question, Q is a question in a question-answer pair. *Cor* is a whole corpus, λ is smoothing parameter. $tf(w,Q)$ is the frequency of word w in a question Q. $|Q|$ and $|Cor|$ denote the length of Q and *Cor*. Jeon proposed a translation-based model (*TM*) to solve the lexical gap problem in the language model [1].

$$P_{TM}(q\,|\,Q) = \prod_{w \in q} (1-\lambda)P_{tr}(w\,|\,Q) + \lambda P_{ml}(w\,|\,Cor) \qquad (2)$$

$$P_{tr}(w\,|\,Q) = \sum_{t \in Q} P(w\,|\,t)P_{ml}(t\,|\,Q), \quad P_{ml}(t\,|\,Q) = \frac{tf(t,Q)}{|Q|},$$

where $P(w\,|\,t)$ denotes the translation probability from word t to word w. Xue proposed a translation-based language model (*TRLM*) by combining the language model and the translation-based model [3].

$$P_{TRML}(q\,|\,Q) = \prod_{w \in q} (1-\lambda)P_{mx}(w\,|\,Q) + \lambda P_{ml}(w\,|\,Cor) \qquad (3)$$

$$P_{mx}(w\,|\,Q) = \delta \sum_{t \in Q} P(w\,|\,t)P_{ml}(t\,|\,Q) + (1-\delta)P_{ml}(w\,|\,Q),$$

Experimental results show that the *TRLM* obtains state-of-the-art performance for the question retrieval. Many researchers consider that three retrieval models, such as *LM*, *TM* and *TRLM*, are baseline retrieval models. We also use these models as baseline retrieval models.

3. Proposed Method

3.1 Extracting Dependency Bigrams from the Dependency Parser

A bigram is a sequence of two adjacent words in a sentence or a document. It reflects the context information. If the window size is small, it does not extract any associated words that are located far away in a sentence. In Fig. 1, although two words (*described_7* and *Jarrett,_1*) are highly relevant, the bigram *described_Jarrett* is not extracted based on the sliding window (window size is less than 6). And, if the window size is increased to extract two associated words remotely located, it can include a lot of noisy bigrams. Therefore, we extract bigrams by using dependencies, called the dependency bigram. We first describe the dependencies based on the graph. In Fig. 1, it is the graph of dependencies in the C&C parser. *described_7* is a main verb and *Jarrett,_1* is the subject. *_7* and *_1* indicate an index in a sentence.

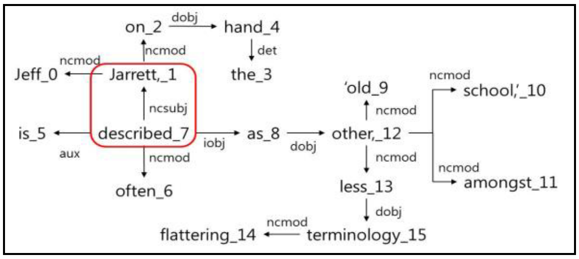

Fig. 1 The graph of dependencies in the C&C parser about a sentence, “Jeff Jarrett, on the hand, is often described as ‘old school,’ amongst other, less flattering terminology.”

In our dependency based model, we just use a head-dependent pair in the dependency as a bigram feature. For example, we extract a head, *described_7*, and its dependent, *Jarrett,_1*. Then we remove an index and generate the dependency bigram (*db*) by combining the head and its dependent, *described_Jarrett,*.

3.2 Dependency Based Model (DM)

The traditional language model is transformed by using the dependency bigram as bigram features instead of unigram features (words) in Eq. (1).

$$P_{DM}(q\,|\,Q) = \prod_{db \in q} (1-\gamma)P_{ml}(db\,|\,Q) + \lambda P_{ml}(db\,|\,Cor) \qquad (4)$$

$$P_{ml}(db\,|\,Q) = \frac{tf(db, Q_{db})}{|Q_{db}|}, \quad P_{ml}(db\,|\,Cor) = \frac{tf(db, Cor_{db})}{|Cor_{db}|},$$

where db indicates the dependency bigram. Q_{db} is dependency bigrams of a question in a question-answer pair. Cor_{db} is dependency bigrams of whole corpus. $|Q_{db}|$ and $|Cor_{db}|$ denote the length of Q_{db} and $|Cor_{db}|$. The distribution of the dependency bigrams are estimated in a question and a corpus based on the maximum likelihood.

The proposed dependency based model is combined with three baselines, *LM*, *TM* and *TRLM*, by the linear combination.

$$P_{DLM}(q\,|\,Q) = \eta P_{DM}(q\,|\,Q) + (1-\eta)P_{LM}(q\,|\,Q) \qquad (5)$$

$$P_{DTM}(q\,|\,Q) = \xi P_{DM}(q\,|\,Q) + (1-\xi)P_{TM}(q\,|\,Q) \qquad (6)$$

$$P_{DTRLM}(q\,|\,Q) = \mu P_{DM}(q\,|\,Q) + (1-\mu)P_{TRML}(q\,|\,Q) \qquad (7)$$

4. Experiments

4.1 Data Sets and Experimental Settings

We used question-answer pairs in the language data of the *Yahoo Webscope Program* (http://webscope.sandbox.yahoo.com/#datasets) for the evaluation. All datasets have been reviewed to conform to Yahoo’s data protection standards, including strict controls on privacy. The datasets are only available for academic use by faculty and university researchers who agree to the Data Sharing Agreement. The dataset *L16* is a small sample of Yahoo! Answers question/answers pages visited following search engine queries in August 2010. It contains 458 queries, 1,571 <query, question, answer> triples, and user’s rating score, which is average user rating of query-question match and query-answer satisfaction. The dataset *L4* (144,067) and question-answer pairs (1,571) in triples of the *L16* are resulting question repository for question retrieval. And we considered

Table 1 The number of queries, target questions, and average answer.

	Query	*Parameter tuning*	*Target*	*Average answer count*
Yahoo	371	20	144,067	2.498

Table 2 Performances of three baseline retrieval models.

	LM	*TM*	*TRLM*
MAP	0.5440	0.6403	**0.6476**

that triples in the *L16* (average user rating of query-question match is less than 2.0)[†] were relevant. To learn the translation probabilities, we used a dataset *L6-1* containing about 2.4 million question-answer pairs in English. Table 1 shows the number of queries, target questions, and average answer count. We selected 20 queries from test data for parameter tuning. Final test query is changed from 391 to 371.

We evaluated the performance of our approach using Mean Average Precision (*MAP*) because there is no need to choose k and ranking mistakes at the top of the ranking are more influential. There are many parameters in our experiments. For the parameters, we did experiments on a small development set of 20 queries to determine the best values. This set was also selected from the Yahoo! Answer randomly and it was not included in the test query set. As a result, we set the smoothing parameter λ in Eqs. (1) & (2) to 0.1, the parameter δ in Eq. (3) to 0.8, and the parameter γ in Eq. (4) to 0.9 through experiments. Then three parameters, such as η, ξ, and μ, for the combination in Eqs. (5), (6) and (7) set 0.8, respectively.

4.2 Performances of Baselines

We tested three baselines, which are the unigram (word) models, as follows:

We compare baselines based on 144,067 of target dataset called *ALL*. As shown in Table 2, *TRLM* achieved better performance than other baselines. We used a pooling strategy to estimate the translation probabilities [3].

4.3 Experimental Results of the Proposed Retrieval Model

We evaluated our proposed models, which are the bigram model (dependency bigram). First, we observed the performance changes of the *DM* according to the different values of the parameter gamma γ in Eq. (4) from 0.1 to 0.9.

DM archived 0.3478. We think that the bigram feature (dependency bigram) works well in the question retrieval. However, since *DM* does not solve the lexical gap problem, the performance was lower than *TM* and *TRLM*. Second, we evaluated combined models in Eqs. (5), (6) and (7). They called the dependency based *LM* (*DLM*), the dependency based *TM* (*DTM*) and the dependency based *TRLM* (*DTRLM*).

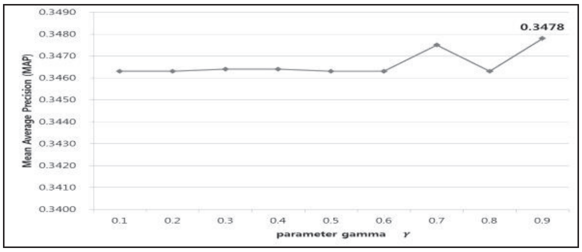

Fig. 2 Performance changes according to the different values by parameter gamma γ in the *DM*.

Table 3 Performances of three combined models[††].

	TRLM	*DLM*	*DTM*	*DTRLM*
50k	0.6874	0.6244 (-9.16%)	**0.7256*** **(+5.56%)**	0.7167* (+4.26%)
100k	0.6685	0.5902 (-11.71%)	**0.6959*** **(+4.10%)**	0.6899* (+3.20%)
ALL	0.6476	0.5736 (-11.43%)	**0.6776*** **(+4.63%)**	0.6703* (+3.51%)

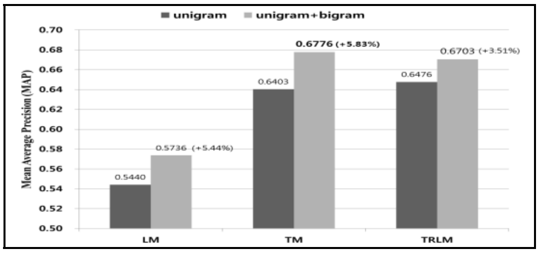

Fig. 3 The performance comparison of all models based on the feature.

As shown in Table 3, the two proposed combined models achieved much better performances than *TRLM*. The models showed relatively 4.63% and 3.51% higher performances than *TRLM* on *MAP*, respectively. We think that the proposed dependency bigram is a very effective feature for the question retrieval.

5. Conclusions and Future Work

This paper has studied the dependency based model using the dependency bigram for the question retrieval. We have constructed a series of experiments with dependency bigram. As a result, when the unigram models (*TM* and *TRLM*) and the bigram model (*DM*) are effectively combined, it achieved the *MAP* of 0.6776 and 0.6703. It is 4.63% and 3.51% relatively higher performance than *TRLM*. We evaluated and analyzed usefulness of dependency bigram features extracted from a dependency parser through experiments and we found that the bigram features are very effective in the question retrieval. In the future, we plan to use word embeddings in a continuous space for the question retrieval.

[†] Each user's rating on query-question match (1 indicating "well matched", 2 indicating "partially matched", 3 indicating "not matched"). We consider two cases (well matched and partially matched) as relevant.

[††] * indicate statistically significant improvements ($p < 0.05$ using a paired t-test) over *TRLM*). *50k, 100k*, and *ALL* are the number of target dataset, such as 50,000, 100,000, and 144,067, respectively.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2015R1D1A1A01056907).

References

[1] J. Jeon, W.B. Croft, and J.H. Lee, "Finding similar questions in large question and answer archives," CIKM '05, pp.84–90, 2005.

[2] H. Duan, Y. Cao, C.Y. Lin, and Y. Yu. "Searching questions by identifying questions topics and question focus," ACL '08, pp.156–164, 2008.

[3] X. Xue, J. Jeon, and W.B. Croft, "Retrieval models for question and answer archives," SIGIR '08, pp.475–482, 2008.

[4] G. Zhou, L. Cai, J. Zhao, and K. Liu, "Phrase-based translation model for question retrieval in community question answer archives," ACL '11, pp.653–662, 2011.

[5] G. Salton, A. Wong, and C.S. Yang, "A vector space model for automatic indexing," Commun. ACM, vol.18, no.11, pp.613–620, 1975.

[6] S. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. "Okapi at trec-3," TREC-3, pp.109–126, 1994.

[7] J.M. Ponte and W.B. Croft, "A language modeling approach to information retrieval," SIGIR '98, pp.275–281, 1998.

[8] P.F. Brown, V.J.D. Pietra, S.A.D. Pietra, and R.L. Mercer, "The mathematics of statistical machine translation: Parameter estimation," Computaional Linguistics, vol.19, no.2, pp.263–311, 1993.

[9] A. Berger and J. Lafferty, "Information retrieval as statistical translation," SIGIR '99, pp.222–229, 1999.

[10] X. Cao, G. Cong, B. Cui, C.S. Jensen, and C. Zhang, "The use of categorization information in language models for question retrieval," CIKM '09, pp.265–274, 2009.

[11] X. Cao, G. Cong, B. Cui, and C.S. Jensen, "A generalized framework of exploring category information for question retrieval in community question answer archives," WWW '10, pp.201–210, 2010.

[12] L. Cai, G. Zhou, K. Liu, and J. Zhao, "Learning the latent topics for question retrieval in community QA," ACL '12, pp.273–281, 2011.

[13] Z. Ji, F. Xu, B. Wang, and B. He, "Question-answer topic model for question retrieval in community question answering," CIKM '12, pp.2471–2474, 2012.

[14] K. Zhang, W. Wu, H. Wu, Z. Li, and M. Zhou, "Question retrieval with high quality answers in community question answering," CIKM '14, pp.371–380, 2014.

[15] C. Zhai and J. Lafferty, "A study of smoothing methods for language models applied to information retrieval," ACM Trans. Information System, vol.22, no.2, pp.179–214, 2004.