HAL
open science

A Word Embedding based Method for Question Retrieval in Community Question Answering

Nouha Othman, Rim Faiz, Kamel Smaili

To cite this version:

Nouha Othman, Rim Faiz, Kamel Smaili. A Word Embedding based Method for Question Retrieval in Community Question Answering. ICNLSSP 2017 - International Conference on Natural Language, Signal and Speech Processing, ISGA, Dec 2017, Casablanca, Morocco. hal-01660005

HAL Id: hal-01660005

https://inria.hal.science/hal-01660005v1

Submitted on 9 Dec 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A Word Embedding based Method for Question Retrieval in Community Question Answering

Nouha Othman[1], *Rim Faiz*[2], *Kamel Smaili*[3],

[1] LARODEC, Institut Supérieur de Gestion de Tunis, Université de Tunis, Tunisia
[2] LARODEC, IHEC Carthage, Université de Carthage, Tunisia
[3]SMarT, LORIA Campus Scientifique BP 139, 54500 Vandoeuvre Lès-Nancy Cedex, France
othmannouha@gmail.com, rim.faiz@ihec.rnu.tn, smaili@loria.fr

Abstract

Community Question Answering (cQA) continues to gain momentum owing to the unceasing rise of user-generated content that dominates the web. CQA are platforms that enable people with different backgrounds to share knowledge by freely asking and answering each other. In this paper, we focus on question retrieval which is deemed to be a key task in cQA. It aims at finding similar archived questions given a new query, assuming that the answers to the similar questions should also answer the new one. This is known to be a challenging task due to the verboseness in natural language and the word mismatch between the questions. Most traditional methods measure the similarity between questions based on the bag-of-words (BOWs) representation capturing no semantics between words. In this paper, we rely on word representation to capture the words semantic information in language vector space. Questions are then ranked using cosine similarity based on the vector-based word representation for each question. Experiments conducted on large-scale cQA data show that our method gives promising results.

Index Terms: Community question answering, Question retrieval, Word embeddings, Cosine similarity

1. Introduction

Over the last decade, with the boom of Web 2.0, the world has witnessed a huge spread of user-generated content, which became a crucial source of information on internet. This brings great attention to the emerging concept of community Question Answering (cQA) that refers to platforms that enable users to interact and answer to other users' questions [10]. Nowadays, there exists a full panoply of cQA services such as Yahoo! Answers[1], Stackoverflow[2], MathOverflow[3] and LinuxQuestions[4]. Such community services have built up massive archives of question-answer pairs that are considered as valuable resources for different tasks like question-answering [21]. The cQA archives are continuously increasing accumulating duplicated questions. As a matter of fact, users cannot easily find the good answers and consequently post new questions that already exist in the archives. In order to avoid wasting time waiting for a new answer, cQA should automatically search the community archive to verify if similar questions have previously been posted. If a similar question is found, its associated answer can be directly returned. Owing to its importance, significant research efforts have been recently put to retrieve similar questions in cQA [21, 3, 2, 16, 22, 12]. Indeed, question retrieval is a non trivial task presenting several challenges, mainly the data sparseness, as questions in cQA are usually very short. Another great challenge is the lexical gap between the queried questions and the existing ones in the archives [21], which constitutes a real obstacle to traditional Information Retrieval (IR) models since users can formulate the same question employing different wording. For instance, the questions: *How to lose weight within a few weeks?* and *What is the best way to get slim fast?*, have the same meaning but they are lexically different. The word mismatching is a critical issue in cQA since questions are relatively short and similar ones usually have sparse representations with little word overlap. From this, it is clear that effective retrieval models for question retrieval are strongly needed to take full advantage of the sizeable community archives.

In order to bridge the lexical gap problem in cQA, most state-of-the-art studies attempt to improve the similarity measure between questions while it is hard to set a compelling similarity function for sparse and discrete representations of words. More importantly, most existing approaches neither take into account the contextual information nor capture enough semantic relations between words. Recently, novel methods for learning distributed word representations, also called word embeddings, have shown significant performance in several IR and Natural Language Processing (NLP) tasks, amongst other questions retrieval in cQA [28]. Word embeddings are low-dimensional vector representations of vocabulary words that capture semantic relationships between them. It is worth noting that to date, the investigation of word embeddings in question retrieval is still in its infancy but the studies in this line are encouraging.

Motivated by the recent success of these emerging methods, in this paper, we propose a word embedding-based method for question retrieval in cQA, *WECOSim*. Instead of representing questions as a bag of words (BoW), we suggest representing them as Bag of-Embedded-Words (BoEW) in a continuous space using word2vec, the most popular word embedding model. Questions are therefore ranked using cosine similarity based on the vector-based word representation for each question. A previous posted question is considered to be semantically similar to a queried question if their corresponding vector representations lie close to each other according to the cosine similarity measure. The previous question with the highest cosine similarity score will be returned as the most similar question to the new posted one. We test the proposed method on a large-scale real data from Yahoo! Answers. Experimental

[1]http://answers.yahoo.com/
[2]http://stackoverflow.com/
[3]http://www.mathoverflow.net
[4]http://www.linuxquestions.org/

results show that our method is promising and can outperform certain state-of-the-art methods for question retrieval in cQA.

The remainder of this paper is organized as follows: In Section (2), we give an overview of the main related work on question retrieval in cQA. Then, we present in Section (3) our proposed word embedding based-method for question retrieval. Section (4) presents our experimental evaluation and Section (5) concludes our paper and outlines some perspectives.

2. Related Work

In cQA, the precision of the returned questions is crucial to ensure high quality answers. The question retrieval task is highly complex due to the lexical gap problem since the queried question and the archived ones often share very few common words or phrases.

Over the recent years, a whole host of methods have been proposed to improve question retrieval in cQA. Several works were based on the vector space model referred to as VSM to calculate the cosine similarity between a query and archived questions [7, 3]. However, the major limitation of VSM is that it favors short questions, while cQA services can handle a wide range of questions not limited to concise or factoid questions. In order to overcome the shortcoming of VSM, BM25 have been employed for question retrieval to take into consideration the question length [3]. Okapi BM25 is the most widely applied model among a family of Okapi retrieval models proposed by Robertson et al. in [15] and has proven significant performance in several IR tasks. Besides, Language Models (LM)s [4] have been also used to explicitly model queries as sequences of query terms instead of sets of terms. LMs estimate the relative likelihood for each possible successor term taking into consideration relative positions of terms. Nonetheless, such models might not be effective when there are few common words between the user's query and the archived questions.

To overcome the vocabulary mismatch problem faced by LMs, the translation model was used to learn correlation between words based on parallel corpora and it has obtained significant performance for question retrieval. The basic intuition behind translation models is to consider question-answer pairs as parallel texts, then relationship of words can be constructed by learning word-to-word translation probabilities such as in [21, 2]. Within the same context, [1] presented a parallel dataset for training statistical word translation models, composed of the definitions and glosses provided for the same term by different lexical semantic resources. In [24], the authors tried to improve the word-based translation model by adding some contextual information when building the translation of phrases as a whole, instead of translating separate words. In [16], the word-based translation model was extended by incorporating semantic information (entities) and explored strategies to learn the translation probabilities between words and concepts using the cQA archives and an entity catalog. Although, the aforementioned basic models have yielded good results, questions and answers are not really parallel, rather they are different from the information they contain [22].

Advanced semantic based approaches were required to further tackle the lexical gap problem and to push the question retrieval task in cQA to the next level. For instance, there were few attempts that have exploited the available category information for question retrieval like in [4, 3, 27]. Despite the fact that these attempts have proven to significantly improve the performance of the language model for question retrieval, the use of category information was restricted to the language model. Wang et al [20] used a parser to build syntactic trees of questions, and rank them based on the similarity between their syntactic trees and that of the query question. Nevertheless, such an approach is very complex since it requires a lot of training data. As observed by [20], existing parsers are still not well-trained to parse informally written questions.

Other works model the semantic relationship between the searched questions and the candidate ones with deep question analysis such as [7] who proposed to identify the question topic and focus for question retrieval. Within this context, some studies relied on a learning-to-ranking strategy like [17] who presented an approach to rank the retrieved questions with multiple features, while [19] rank the candidate answers with a single word information instead of the combination of various features. Latent Semantic Indexing (LSI) [6] was also employed to address the given task like in [14]. While being effective to address the synonymy and polysemy by mapping words about the same concept next to each other, the efficiency of LSI highly depends on the data structure.

Otherwise, other works focused on the representation learning for questions, relying on an emerging model for learning distributed representations of words in a low-dimensional vector space namely Word Embedding. This latter has recently been subject of a wide interest and has shown promise in numerous NLP tasks [18, 5], in particular for question retrieval [28]. The main virtue of this unsupervised learning model is that it doesn't need expensive annotation; it only requires a huge amount of raw textual data in its training phase. As we believe that the representation of words is vital for the question retrieval task and inspired by the success of the latter model, we rely on word embeddings to improve the question retrieval task in cQA.

3. Description of WECOSim

The intuition behind the method we propose for question retrieval, called *WECOSim*, is to transform words in each question in the community collection into continuous vectors. Unlike traditional methods which represent each question as Bag Of Words (BOWs), we propose to represent a question as a Bag-of-Embedded-Words (BoEW). The continuous word representations are learned in advance using the continuous bag-of-words (CBOW) model [11]. Each question is, therefore, be defined as a set of words embedded in a continuous space. Besides, the cosine similarity is used to calculate the similarity between the average of the word vectors corresponding to the queried question and that of each existing question in the archive. The historical questions are then ranked according to their cosine similarity scores in order to return the top ranking question having the maximum score, as the most relevant one to the new queried question. The proposed method for question retrieval in cQA consists of three steps namely, question preprocessing, word embedding learning and question ranking.

3.1. Question Preprocessing

The question preprocessing module intends to process the natural language questions and extract the useful terms in order to generate formal queries. These latter are obtained by applying text cleaning, tokenization, stopwords removal and stemming. Thus, at the end of the question preprocessing module, we obtain a set of filtered queries, each of which is formally defined as follows: $Q = \{t_1, t_2, ..., t_q\}$ where t represents a separate term of the query Q and q denotes the number of query terms.

3.2. Word Embedding Learning

Word embedding techniques, also known as distributed semantic representations play a significant role in building continuous word vectors based on their contexts in a large corpus. They learn a low-dimensional vector for each vocabulary term in which the similarity between the word vectors can show the syntactic and semantic similarities between the corresponding words. Basically, there exist two main types of word embeddings namely Continuous Bag-of-Words model (CBoW) and Skip-gram model. The former one consists in predicting a current word given its context, while the second does the inverse predicting the contextual words given a target word in a sliding window. It is worthwhile to note that, in this work, we consider the CBOW model [11] to learn word embeddings, since it is more efficient and performs better with sizeable data than Skip-gram. As shown in Figure 1, the CBOW model predicts the center word given the representation of its surrounding words using continuous distributed bag-of-words representation of the context, hence the name CBOW. The context vector is got by

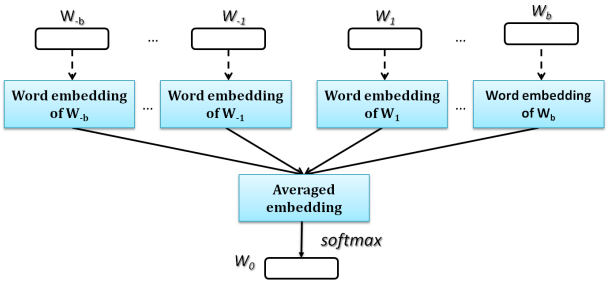

Figure 1: Overview of the Continuous Bag-of-Words model.

averaging the embeddings of each contextual word while the prediction of the center word w_0 is obtained by applying a softmax over the vocabulary V. Formally, let d be the word embedding dimension, the output matrix $O \in \Re^{|V| \times d}$ maps the context vector c into a $|V|$-dimensional vector representing the center word, and maximizes the following probability:

$$p(v_0 \mid w_{[-b,b]-\{0\}}) = \frac{\exp v_0^T O_c}{\sum_{v \in V} \exp v^T O_c} \quad (1)$$

where b is a hyperparameter defining the window of context words, O_c represents the projection of the context vector c into the vocabulary V and v is a one-hot representation. The strength of CBOW is that it does not rise substantially when we increase the window b.

3.3. Question Ranking

Once the questions are presented as Bag of-Embedded-Words (BoEW), we compute the average vector v_q of the queried question. Similarly, for each historical question, we calculate its average vector v_d. The similarity between a queried question and a historical one in the vector space is calculated as the cosine similarity between v_q and v_d.

4. Experiments

4.1. Dataset

In our experiments, we used the dataset released by [23] for evaluation. In order to construct the dataset, the authors crawled questions from all categories in Yahoo! Answers, the most popular cQA platform, and then randomly splitted the questions into two sets while maintaining their distributions in all categories. The first set contains 1,123,034 questions as a question repository for question search, while the second is used as the test set and contains 252 queries and 1624 manually labeled relevant questions. The number of relevant questions related to each original query varies from 2 to 30. The questions are of different lengths varying from two to 15 words, in different structures and belonging to various categories e.g. Computers and Internet, Yahoo! Products, Entertainment and Music, Education and Reference, Business and Finance, Pets, Health, Sports, Travel, Diet and Fitness. Table 1 shows an example of a query and its corresponding related questions from the test set. To train the word embeddings, we resorted to another large-scale data set from cQA sites, namely the Yahoo! Webscope dataset[5], including 1,256,173 questions with 2,512,345 distinct words. Some preprocessing was performed before the experiments; all questions were lower cased, tokenized, stemmed by Porter Stemmer[6] and all stop words were removed.

Table 1: *Example of questions from the test set.*

Query:	How can I get skinnier without getting in a diet?
Category:	Diet and Fitness
Topic:	Weight loss
Related questions	- How do I get fit without changing my diet? - How can i get slim but neither diet nor exercise? - How do you get skinny fast without diet pills? - I need a solution for getting fit (loosing weight) and I must say I cant take tough diets ?

4.2. Learning of Word Embedding

We trained the word embeddings on the whole Yahoo! Webscope dataset using word2vec in order to represent the words of the training data as continuous vectors which capture the contexts of the words. The training parameters of word2vec were set after several tests: the dimensionality of the feature vectors was fixed at 300 (size=300), the size of the context window was set to 10 (window=10) and the number of negative samples was set to 25 (negative=25).

4.3. Evaluation Metrics

In order to evaluate the performance of our method, we used Mean Average Precision (MAP) and Precision@n (P@n) as they are extensively used for evaluating the performance of question retrieval for cQA. Particularly, MAP is the most commonly used metric in the literature assuming that the user is interested in finding many relevant questions for each query. MAP rewards methods that not only return relevant questions early, but also get good ranking of the results. Given a set of queried questions Q, MAP represents the mean of the average precision for each queried question q and it is set as follows: MAP $= \frac{\sum_{q \in Q} \text{AvgP}(q)}{|Q|}$ where $AvgP(q)$ is the mean of the precision scores after each relevant question q is retrieved.

Precision@n returns the proportion of the top-n retrieved questions that are relevant. Given a set of queried questions

[5]The Yahoo! Webscope dataset Yahoo answers comprehensive questions and answers version 1.0.2, available at "http://research.yahoo.com/Academic_Relations"

[6]http://tartarus.org/martin/PorterStemmer/

Q, P@n is the proportion of the top n retrieved questions that are relevant to the queries, and it is defined as follows: $\text{P@n} = \frac{1}{|Q|}\sum_{q \in Q}\frac{\text{Nr}}{N}$ where Nr is the number of relevant questions among the top N ranked list returned for a query q. In our experiments, we calculated P@10 and P@5.

4.4. Main Results

We compare the performance of WECOSim with the following competitive state-of-the-art question retrieval models tested by Zhang et al. in [23] on the same dataset:

- **TLM** [21]: A translation based language model which combines the translation model estimated using the question and the language model estimated using the answer part. It integrates word-to-word translation probabilities learned by exploiting various sources of information.
- **PBTM** [24]: A phrase based translation model which employs machine translation probabilities and assumes that question retrieval should be performed at the phrase level. TLM learns the probability of translating a sequence of words in a historical question into another sequence of words in a queried question.
- **ETLM** [16]: An entity based translation language model, which is an extension of TLM by replacing the word translation with entity translation in order to incorporate semantic information within the entities.
- **WKM** [29]: A world knowledge based model which used Wikipedia as an external resource to add the estimation of the term weights to the ranking function. A concept thesaurus was built based on the semantic relations extracted from the world knowledge of Wikipedia.
- **M-NET** [28]: A continuous word embedding based model, which integrates the category information of the questions to get the updated word embedding, assuming that the representations of words that belong to the same category should be close to each other.
- **ParaKCM** [23]: A key concept paraphrasing based approach which explores the translations of pivot languages and expands queries with the paraphrases. It assumes that paraphrases contributes additional semantic connection between the key concepts in the queried question and those of the historical questions.

From Table 2, we can see that PBTM outperforms TLM which demonstrates that capturing contextual information in modeling the translation of phrases as a whole or consecutive sequence of words is more effective than translating single words in isolation. This is because, by and large, there is a dependency between adjacent words in a phrase. The fact that ETLM (an extension of TLM) performs as good as PBTM proves that replacing the word translation by entity translation for ranking improves the performance of the translation language model.

Table 2: *Comparison of the question retrieval performance of different models.*

	TLM	PBTM	ETLM	WKM	M-NET	ParaKCM	WECOSim
P@5	0.3238	0.3318	0.3314	0.3413	0.3686	0.3722	**0.3432**
P@10	0.2548	0.2603	0.2603	0.2715	0.2848	0.2889	**0.2738**
MAP	0.3957	0.4095	0.4073	0.4116	0.4507	0.4578	**0.4125**

Although, ETLM and WKM are both based on external knowledge resource e.g. Wikipedia, WKM uses wider information from the knowledge source. Specifically, WKM builds a Wikipedia thesaurus, which derives the concept relationships (e.g. synonymy, hypernymy, polysemy and associative relations) based on the structural knowledge in Wikipedia. The different relations in the thesaurus are treated according to their importance to expand the query and then enhance the traditional similarity measure for question retrieval. Nevertheless, the performance of WKM and ETLM are limited by the low coverage of the concepts of Wikipedia on the various users' questions. The results show that our method WECOSim slightly outperforms the aforementioned methods by returning a good number of relevant questions among the retrieved ones early. A possible reason behind this is that context-vector representations learned by word2vec can effectively address the word lexical gap problem by capturing semantic relations between words, while the other methods do not capture enough information about semantic equivalence. We can say that questions represented by bag-of-embedded words can be captured more accurately than traditional bag-of-words models which cannot capture neither semantics nor positions in text. This good performance indicates that the use of word embeddings along with cosine similarity is effective in the question retrieval task. However, we find that sometimes, our method fails to retrieve similar questions when questions contain misspelled query terms. For instance, questions containing *sofwar* by mistake cannot be retrieved for a query containing the term *software*. Such cases show that our approach fails to address some lexical disagreement problems. Furthermore, there are few cases where WECOSim fails to detect semantic equivalence. Some of these cases include questions having one single similar question and most words of this latter do not appear in a similar context with those of the queried question. M-NET, also based on continuous word embeddings performs better than our method owing to the use of metadata of category information to encode the properties of words, from which similar words can be grouped according to their categories. The best performance is achieved by ParaKCM, a key concept paraphrasing based approach which explores the translations of pivot languages and expands queries with the generated paraphrases for question retrieval.

5. Conclusion

In this paper, we lay out a word embedding based method to tackle the lexical gap problem in question retrieval from cQA archives. In order to find semantically similar questions to a new query, previous posted questions are ranked using cosine similarity based on their vector-based word representations in a continuous space. Experimental results conducted on large-scale cQA data show the effectiveness of the proposed method. However, word embedding models assume that each word preserves only a single vector. It is the reason why it faces lexical ambiguity due to polysemy and homonymy, and it is therefore an important problem to address. On the other hand, while the cosine similarity is shown to be effective in identifying semantically closest words, this measure becomes insufficient when the order of words is not needed. In future work, we look forward to improving our method by investigating the performance of certain powerful techniques such as Latent Semantic Indexing (LSI) along with word embeddings. We also consider incorporating various types of metadata information into the learning process in order to enrich word representations.

6. References

[1] Bernhard, D. and Gurevych, I., "Combining lexical semantic resources with question and answer archives for translation-based answer finding", In Proceedings of ACL, pages 728–736, 2009.

[2] Cai, L., Zhou, G., Liu, K., and Zhao, J., "Learning the latent topics for question retrieval in community qa", In Proceedings of IJCNLP, pages 273–281, 2011.

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

[4] Cao, X., Cong, G., Cui, B., Jensen, C. S., and Zhang, C., "The use of categorization information in language models for question retrieval", In Proceedings of the 18th ACM conference on Information and knowledge management, pages 265–274, 2009.

[5] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P., "Natural language processing (almost) from scratch", Journal of Machine Learning Research, pages 2493–2537, 2011.

[6] Deerwester, S., Dumais, S. T., Furnas, G.W., Landauer, T. K., and Harshman, R., "Indexing by latent semantic analysis", Journal of the American society for information science, 41(6):391, 1990.

[7] Duan, H., Cao, Y., Lin, C.-Y., and Yu, Y., "Searching questions by identifying question topic and question focus", s. In Proceedings of 46th Annual Meeting of the Association for Computational Linguistics: Human Language Tchnologies (ACL:HLT), volume 8, pages 156–164, 2008.

[8] Kenter, T. and De Rijke, M., "Short text similarity with word embeddings", In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, pages 1411–1420, 2015.

[9] Levy, O., Goldberg, Y., and Dagan, I., "Improving distributional similarity with lessons learned from word embeddings", Transactions of the Association for Computational Linguistics, 3:211–225, 2015.

[10] Liu, Y., Bian, J., and Agichtein, E., "Predicting information seeker satisfaction in community question answering", In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pages 483–490, 2008.

[11] Mikolov, T., Chen, K., Corrado, G., and Dean, J., "Efficient estimation of word representations in vector space", arXiv preprint arXiv:1301–3781, 2013.

[12] Nakov, P., Hoogeveen, D., M'arquez, L., Moschitti, A., Mubarak, H., Baldwin, T., and Verspoor, K., "Semeval-2017 task 3: Community question answering", In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval–2017), pages 27–48, 2017.

[13] Othman, N. and Faiz, R.,"A multilingual approach to improve passage retrieval for automatic question answering", In International Conference on Applications of Natural Language to Information Systems, Springer, pages 127–139, 2016.

[14] Qiu, X., Tian, L., and Huang, X., "Latent semantic tensor indexing for community-based question answering", In In Proc. the 51st Annual Meeting of the Association for Computational Linguistics, pages 434–439, 2013.

[15] Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M., Gatford, M., et al., Okapi at TREC-3, Nist Special Publication Sp, 109:109, 1995.

[16] Singh, A., "Entity based q&a retrieval", In Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pages 1266–1277, 2012.

[17] Surdeanu, M., Ciaramita, M., and Zaragoza, H., "Learning to rank answers on large online qa collections", . In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT), volume 8, pages 719–727, 2008.

[18] Turian, J., Ratinov, L., and Bengio, Y., "Word representations: a simple and general method for semisupervised learning", In Proceedings of the 48th annual meeting of the association for computational linguistics, pages 384–394, 2010.

[19] Wang, B., Wang, X., Sun, C., Liu, B., and Sun, L., "Modeling semantic relevance for question-answer pairs in web social communities", In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 1230–1238, 2010.

[20] Wang, K., Ming, Z., and Chua, T.-S., "A syntactic tree matching approach to finding similar questions in community-based qa services", In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, pages 187–194, 2009.

[21] Xue, X., Jeon, J., and Croft, W. B., "Retrieval models for question and answer archives", In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pages 475–482, 2008.

[22] Zhang, K., Wu, W., Wu, H., Li, Z., and Zhou, M., "Question retrieval with high quality answers in community question answering". In Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, pages 371–380, 2014.

[23] Zhang, W.-N., Ming, Z.-Y., Zhang, Y., Liu, T., and Chua, T.-S., "Capturing the semantics of key phrases using multiple languages for question retrieval", IEEE Transactions on Knowledge and Data Engineering, 28(4):888–900, 2016

[24] Zhou, G., Cai, L., Zhao, J., and Liu, K., "Phrase-based translation model for question retrieval in community question answer archives", In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, pages 653–662, 2011.

[27] Zhou, G., Chen, Y., Zeng, D., and Zhao, J., "Towards faster and better retrieval models for question search", In Proceedings of the 22nd ACM international conference on Conference on information and knowledge management, pages 2139–2148, 2013.

[28] Zhou, G., He, T., Zhao, J., and Hu, P., "Learning continuous word embedding with metadata for question retrieval in community question answering", In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, pages 250–259, 2015.

[29] Zhou, G., Liu, Y., Liu, F., Zeng, D., and Zhao, J. (2013b). "Improving question retrieval in community question answering using world knowledge", In IJCAI, volume 13, pages 2239–245, 2013.