Modeling and Learning Continuous Word Embedding with Metadata for Question Retrieval

C.Hari[1], G.V.Ramesh Babu[2]
[1] *Student. Dept. of Computer Science , SVU College of CM&CS,Tirupati*
[2] *Assistant Professor, Dept. of Computer Science ,SVU College of CM&CS,Tirupati*

Abstract- **Community question answering (cQA) has become an important issue due to the popularity of cQA archives on the Web. This paper focuses on addressing the lexical gap problem in question retrieval. Question retrieval in cQA archives aims to find the existing questions that are semantically equivalent or relevant to the queried questions. However, the lexical gap problem brings new challenge for question retrieval in cQA. In this paper, we propose to model and learn continuous word embeddings with metadata of category information within cQA pages for question retrieval using two novel category powered models. One is basic category powered model called MB-NET and the other one is enhanced category powered model called ME-NET which can better learn the word embeddings and alleviate the lexical gap problem. To deal with the variable size of word embedding vectors, we employ the framework of fisher kernel to aggregate them into the fixed-length vectors. Experimental results on large-scale English and Chinese cQA data sets show that our proposed approaches can significantly outperform state-of-the-art translation models and topic-based models for question retrieval in cQA. Moreover, we further conduct our approaches on large-scale automatic evaluation experiments. The evaluation results show that promising and significant performance improvements can be achieved.**

1. INTRODUCTION

OVER the past few years, a large amount of user-generated content has become an important information resource on the Web. These include the traditional Frequently Asked Questions(FAQ) archives and the emerging community question answering (cQA) services, such as Yahoo! Answers1, Live QnA2, and Baidu Zhidao3. The content in these web sites is usually organized as questions and lists of answers associated with metadata like user chosen categories to questions and askers' awards to the best answers. This data made cQA archives valuable resources for various tasks like question-answering [1], [2] and knowledge mining [3], etc.

One fundamental task for reusing content in cQA is to find similar questions for queried questions, as questions are the keys to accessing the knowledge in cQA. Then the best answers of these similar questions will be used to answer the queried questions. This is what we call question retrieval in this article. Compared to the traditional ad hoc information retrieval, question retrieval in cQA has several advantages. First, the user can use natural language instead of only keywords as a query, and thus can poten-tially express his or her information need more clearly. Second, the system returns several possible answers directly instead of a long list of ranked documents, and can therefore increase the efficiency of finding the required answers [2].

in cQA is how to deal with the lexical gap between the queried questions and the existing questions in the archives. Lexical gap means that the queried questions may contain words that are different from, but related to, the words in the existing questions. For example shown in [10], we find that for a queried question "how do I get knots out of my cats fur?", there are good answers under an existing question "how can I remove a tangle in my cat's fur?" in Yahoo! Answers. Although the two questions share few words in common, they have very similar meanings, it is hard for traditional retrieval models (e.g., BM25 [11]) to determine their similarity. This lexical gap has become a major barricade preventing traditional IR models (e.g., BM25) from retrieving similar questions in cQA.

To address the lexical gap problem in cQA, previous work in the literature can be divided into two groups. The first group is the translation models, which leverage the question-answer pairs to learn the semantically related words for improving traditional

IR models [1], [2], [8]. The basic assumption is that question-answer pairs are "parallel texts" and relationship of words (or phrases) can be established through word-to-word (or phrase-to-phrase) translation probabilities [1], [2], [8]. Experimental results show that translation models obtain state-of-the-art performance for question retrieval in cQA. However, questions and answers are far from "parallel" in practice, questions and answers are highly asymmetric on the information they contain [10]. The second group is the topic-based models [12], [13], which learn the latent topics aligned across the question-answer pairs to alleviate the lexical gap problem, with the assumption that a question and its paired answers share the same topic distribution. However, questions and answers are heterogeneous in many aspects, they do not share the same topic distribution in practice.

Inspired by the recent success of continuous space word representations in capturing the semantic similarities in various natural language processing tasks, we propose to incorporate an embedding of words in a continuous space for question repre-sentations. Recently, some efforts, such as skip-gram model [14], DEEPMATCH architecture [15], convolutional neural tensor net-work architecture (CNTN) [16], have attempted to learn word embeddings that can capture the semantic information among natural language questions, or learn question representations with deep neural networks. However, these existing methods still lack the capability of encoding the properties of words and the complex relationships among words very well, since questions in cQA often contains incomplete and ambiguous information. Fortunately, the metadata (e.g., category, user ratings, social signals,etc.)

In cQA provide a golden mine for enhancing the quality of learned word embeddings or question representations. In particular, when a user asks a question in cQA site, the user is typically required to choose a category for the question from the predefined categories. The available categories encode the attributes or properties of questions [7], [12], [17], [18], [19]. Therefore, we propose a novel metadata powered framework to leverage the category information to produce the word representations of higher quality. Specially, we propose two models: one is basic category powered model and the other one is enhanced category powered model. We build the regularization function derived from the metadata of category information along with the training process of the skip-gram model. By solving the optimization problem using back propagation neural networks, we can obtain word representations enhanced by the category knowledge.

Once the words are embedded in a continuous space, one can view a question as a Bag-of-Embedded-Words (BoEW). Then, the variable-cardinality BoEW will be aggregated into a fixed-length vector by using the Fisher kernel (FK) framework of [20],[21]. Through the two steps, the proposed approaches can map a question into a length invariable compact vector, which can be efficiently and effectively for large-scale question retrieval task in cQA.

We evaluate the proposed approaches on large-scale Yahoo! Answers data and Baidu Zhidao data. Yahoo! Answers and Baidu Zhidao represent the largest and most popular cQA archives in English and Chinese, respectively. We conduct both quantitative and qualitative evaluations. Experimental results show that our approaches can significantly outperform state-of-the-art transla-tion models and topic-based models for question retrieval in cQA. In summary, we make the following contributions:

We study the task of question retrieval in cQA and rep-resent a question as a bag-of-embedded-words (BoEW) in a continuous space. In order to learn better word embed-dings, we introduce two different models: one is the basic category powered model and the other one is the enhanced category powered model. These word embedding models contribute the most the performance.

We introduce a novel method to aggregate the variable-cardinality BoEW into a fixed-length vector by using the FK. The FK is just one possible way to subsequently transform this bag representation into a fixed-length vector which is more amenable to large-scale processing.

Experiments conducted on English and Chinese cQA data sets demonstrate the effectiveness of our approaches. Experimental results show that our approaches signifi-cantly outperform the state-of-the-art translation models and topic-based models. We also show that metadata of category information benefits the word embedding learn-ing for question

representation. Moreover, the enhanced category powered model can better model the word rep-resentation than the basic category powered model. In addition, we further conduct a series of experiments to evaluate our proposed approaches automatically on large-scale data sets. The experimental results demonstrate that our approaches can significantly outperform state-of-the-art models for question retrieval in cQA.

The remainder of this article is organized as follows. Section 2 summarizes the related work. In Section 3, we describe the pro-posed metadata powered word representation models for question retrieval. Section 4 reports the experimental results. We conclude with ideas for future research in Section 5.

2RELATED WORK

2.1 Question Retrieval in cQA

In recent years, with the flouring of community question an-swering (cQA) archives, significant research efforts have been conducted in attempt to improve question retrieval in cQA [1], [2], [4], [5], [6], [7], [8], [9], [10], [22], [23], [24]. Particularly, language model based methods are proven effective. Most cQA researchers focus on leveraging metadata in cQA to improve the performance of the traditional language models for question retrieval [10], [25]. Basically, there are five groups of work.

The first group considers leveraging categories of questions. For example, Ming et al [18] proposed a framework to seamlessly integrate category-specific term weight into the existing VSM and BM25 retrieval models for question retrieval. Cao et al. [17], [60] employed classifiers to compute the probability of a question be-longing to different categories, and then incorporated the classified categories into language model for question retrieval. Then, Cao et al. [7], [60] introduced the different combinations to compute the global relevance and the local relevance for question retrieval, the combination VSM + WTLM showed the superior performance than others. Furthermore, Ji et al. [61] proposed a category-integrated language model (CLM) for question retrieval, which views category-specific term saliency as the Dirichlet hyper-parameter that weights the parameters of LM. Zhou et al. [19] proposed a faster and better retrieval model by leveraging category to filter certain amount of irrelevant questions under a wide range of leaf categories. Zhou et al. [23] proposed a novel approach called group non-negative matrix factorization with natural cat-egories for question retrieval. This is achieved by learning the category-specific topics for each category as well as shared topics across all categories via a group non-negative matrix factorization framework. Our preliminary study has shown that metadata of category information benefits the word embedding learning for question representation. In our previous study, we proposed a basic category powered model for word embedding learning based on the same leaf category assumption, with potential relevant questions under the similar leaf categories being omitted [24]. However, as we will show, this simple model does not necessarily encode the attributes or properties of words. How to encode the attributes or properties of words from the similar categories still needs to be further investigated. We now extend our preliminary study and propose an enhanced category powered model to better learn the word embeddings for question retrieval. The enhanced category powered model is shown to be more effective. We also conduct more comprehensive comparisons with in-depth analysis and further evaluate our newly proposed approaches on large-scale English and Chinese cQA data sets.

The second group leverages question-answer pairs to learn various translation models to bridge the lexical gap problem. For example, Jeon et al. [1] proposed a word-based translation model which exploits the semantic similarity between answers of existing questions to learn translation probabilities, which allows them to match semantically similar questions despite lexical gap. Xue et al. [2] proposed a word-based translation language model for question retrieval with a query likelihood model for the answer. Experiments consistently reported that the word-based translation model could yield better performance than the traditional meth-ods (e.g., VSM, BM25 and LM). However, these word-based translation models are considered to be context independent in that they do not take into account any contextual information in modeling word translation probabilities. In order to further improve the word-based translation model with some contextual information, Riezler et al. [26] and Zhou et al. [8] proposed a phrase-based translation model for

question and answer retrieval. The phrase-based translation model can capture some contextual information in modeling the translation of phrases as a whole, thus the more accurate translations can better improve the retrieval performance. Furthermore, Singh [9] addressed the lexical gap issues by extending the lexical word-based translation model to incorporate semantic information (entities). However, since it is possible for unimportant words (e.g., non-topical words, common words) to be included in the translation models, a lack of noise control on the models can cause degradation of retrieval performance. Lee et al. [5] investigated a number of empirical methods for eliminating unimportant words in order to construct compact translation models for retrieval purpose. Bernhard and Gurevych [6] proposed to use as a parallel training data set the definitions and glosses provided for the same term by different lexical semantic resources. Besides, Zhou et al. [27] proposed to use of translated words to enrich the question representation, going beyond the words in the original language to represent a question. Zhou et al. [28] proposed to employ statistical machine translation to improve question retrieval and enrich the question representation with the translated words from other languages via matrix factorization. Zhang et al. [29], [30] explored a pivot language translation based approach to derive the paraphrases of key concepts.

The third group applies topic modeling techniques for infor-mation retrieval. In recent years, probabilistic topic models have also been introduced to cQA. For example, Cai et al. [12] proposed a topic model incorporated with the category information into the process of discovering the latent topics in the content of questions. Then they combine the semantic similarity based on latent topics with the translation-based language model [2] into a unified framework for question retrieval. Ji et al. [13] proposed a question-answer topic model to learn the latent topics aligned across the question-answer pairs to alleviate the lexical gap problem, with the assumption that a question and its paired answer share the same topic distribution. Zhang et al. [10] proposed a supervised question-answer topic modeling approach, which assumes that questions and answers share some common latent topics and are generated in a question language and answer language. Besides, other researchers also applied the topic models for the related tasks in cQA. Guo et al. [31] proposed a generative model to simulate user behaviors in cQA, for both question asking and answering, and then simultaneously obtain topic analysis of questions/answers and users. Then they recommended answer providers for new questions according to discovered topic as well as term-level information of questions and users. Zhou et al. [32] proposed a topic-sensitive probabilistic model by taking into account both the link structure and the topical similarity among users for expert finding.

The fourth group employs the syntactic information for ques-tion retrieval. For example, Duan et al. [4] first detected question topic and question focus by using a tree cut method and syntactic parser. They then proposed a new language model to capture the relation between question topic and question focus for question retrieval. After that, Wang et al. [33] proposed a syntactic tree matching model to finding similar questions, and demonstrated that the model is robust against grammatical errors.

The fifth group applies deep learning for question retrieval. Recently, deep learning gains widely interest in natural language processing community. Wang et al. [34] applied the deep belief nets (DBN) to model the relevance of question-answer pairs in cQA, by calculating the distance in the latent semantic space produced by DBN. Hu et al. [35] used a DBN to learn joint representations for textual features and non-textual features, and a linear classification layer on these features is used to predict high quality answers. Lu et al. [15] proposed a DEEPMATCH architecture where patches are extracted to encode low level lexical interaction between question and answer, and a multi-layer regression model calculates the matching score from those patches. Mohit et al. [36] trained a recursive neural network (RNN) based on dependency tree for factoid question answering, by mapping the question into latent space. Hu et al. [37] proposed a deep convolutional architecture for matching natural language sentences (e.g., question and answer), which can nicely combine the hierarchical modeling of individual sentences and the patterns of their matching. Qiu and Huang [16] proposed a convolutional neural tensor network architecture to encode the sentences (e.g., question or answer) in semantic space and model their interactions with a

tensor layer. This model integrated sentence modeling and semantic matching into a single model, which cannot only capture the useful information with convolutional and pooling layers, but also learn the matching metrics between the question and its answer. In contrast to the works described above that assume question-answer pairs are "parallel text", our paper deals with the lexical gap by learning continuous word embeddings in capturing the similarities without any assumptions, which is much more reasonable in practice.

Besides, some other studies model the semantic relationship between queries and answers with deep linguistic analysis or a learning to rank strategy. Surdeanu et al. [38] and Carmel et al. [39] proposed an approach to rank the answers retrieved by Yahoo! Answers with multiple features. Wang et al. [40] aimed to rank the candidate answers with only word information instead of the combination of different kinds of features. Zhou et al. [41] developed a unified framework to leverage the semantic knowledge to enhance the question retrieval in the concept space.

2.2 Word Embedding Learning

Representation of words as continuous vectors has attracted increasing attention in the area of natural language process-ing (NLP). Recently, a series of works applied deep learning techniques to learn high-quality word representations. Bengio et al. [42] proposed a probabilistic neural network language model (NNLM) for word representations. Furthermore, Mikolov et al. [14] proposed efficient neural network models for learning

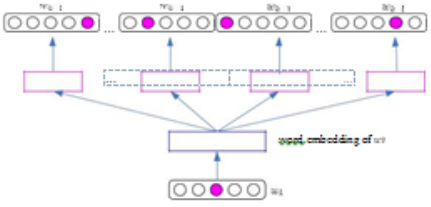


Fig. 1. The skip-gram model, which predicts surrounding words given the current word.

word representations, including the continuous skip-gram model and the continuous bag-of-word model (CBOW), both of which are unsupervised models learned from large-scale text corpora. Besides, there are also a large number of works addressing the task of learning word representations [43], [44], [45].

Nevertheless, since most the existing works learned word rep-resentations mainly based on the word co-occurrence information, the obtained word embeddings cannot capture the relationship between two syntactically or semantically similar words if either of them yields very little context information. On the other hand, even though amount of context could be noisy or biased such that they cannot reflect the inherent relationship between words and further mislead the training process. Most recently, Yu et al. [46] used semantic prior knowledge to improve word representations. Xu et al. [47] used the knowledge graph to advance the learning of word embeddings. In contrast to all the aforementioned works, in this paper, we present a general method to leverage the metadata of category information within cQA pages to further improve the word embedding representations. To our knowledge, it is the first work to learn word embeddings with metadata on cQA data set.

3.OUR APPROACH

In this section, we describe the proposed approach: learning continuous word embedding with metadata for question retrieval in cQA. The proposed framework consists of two steps: (1) word embedding learning step: given a cQA data collection, questions are treated as the basic units. For each word in a question, we firstly transform it to a continuous word vector through the looking up tables. Once the word embeddings are learned, each question is represented by a variable-cardinality word embedding vector (also called BoEW); (2) fisher vector generation step: which uses a generative model in the FK framework to generate fisher vectors (FVs) by aggregating the BoEWs for all the questions. Question retrieval can be performed through calculating the similarity be-tween the FVs of a queried question and an existing question in the archive.

From the framework, we can see that although the word embedding learning computations and generative model estima-tion are time consuming, they can run only once in advance. Meanwhile, the computational requirements of FV generation and similarity calculation are limited. Hence, the proposed framework can efficiently achieve the large-scale question retrieval task.

3.1 Word Embedding Learning

Word embedding learning has gained widely interests for a variety of NLP tasks (e.g., Chinese word

segmentation and part-of-speech tagging (POS) [48], named entity recognition [49], dependency parsing [50], sentiment analysis [51], information extraction [52],question answering [24], [37], etc.). The basic idea is that similar words tend to be close to each other with the vector representation. Mikolov et al. [14] also demonstrate the learned word embedding representations could capture meaningful syntactic and seman-tic regularities. Among the various word embedding learning methods, we consider the context-aware predicting model, more specifically, the Skip-gram model [14] and continuous bag-of-words model (CBOW) [14] for learning word embeddings, since they are much more efficient as well as memory-saving than other approaches.

Let wk represent the kth words in the given words sequence fw1; w2; ; wN g. For easy explanation, we take the example of Skip-gram model to describe the details. In the Skip-gram model (see Fig. 1), a sliding window is employed on the input text stream to generate the training data, and l indicates the context window size to be 2l + 1. In each slide window, the model aims to use the central word wk as input to predict the context words. Let Md N denote the learned embedding matrix, where N is the vocabulary size and d is the dimension of word embeddings. Each column of Mrepresents the embedding of a word. Let wk is first mapped to its embedding ewk by selecting the corresponding column vector of M. The probability of its context word wk+j is then computed using a log-linear softmax function:

$$p(w_{k+j}|w_k;) = \frac{\exp(e_{w}^{T}{}_{k+} e_{w_k})}{\sum_{w=1}^{N} \exp(e_{w}^{T} e_{w_k})}$$

where are the parameters we should learned, k = 1d, and j2 [l; l]. Then, the log-likelihood over the entire training data can be computed as:

$J(X) = \log p(w_{k+j}|w_k;)$

$(w_k;w_{k+j})$

To calculate the prediction errors for back propagation, we need to compute the derivative of p(wk+jjwk;), whose com-putation cost is proportional to the vocabulary size N. As N is often very large, it is difficult to directly compute the derivative. To deal this problem, Mikolov et al. [14] proposed a simple negative sampling method, which generates r noise samples for each input word to estimate the target word, in which r is a very small number compared with N. Therefore, the training time yields linear scale to the number of noise samples and it becomes independent of the vocabulary size. Suppose the frequency of word w is u(w), then the probability of sampling w is usually set to p(w) / u(w)3=4 [14].

Although we only use the skip-gram model to illustrate our approach, the similar framework can be developed on the basis of any other word embedding models.

3.2 Metadata Powered Modeling

After briefing the skip-gram model, we introduce how we equip them with the metadata information.

3.2.1 Basic Category Powered Model

In cQA sites, there are several metadata, such as "cate-gory","voting" and so on. In this paper, we only consider the metadata of category information for word embedding learning. All questions in cQA are usually organized into a hierarchy of categories shown in Fig. 2. When an user asks a question, the user typically required to choose a category label for the question from a predefined hierarchy of categories [7], [19]. Previous work in

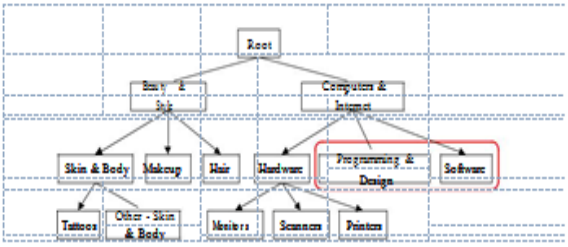


Fig. 2. An example of category hierarchy in Yahoo! Answers.

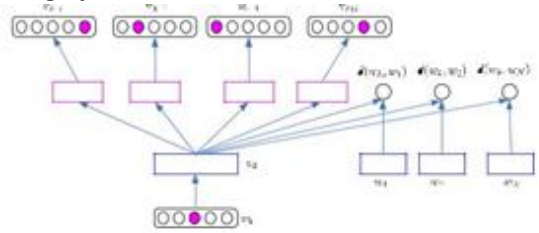
the literature has demonstrated the effectiveness of the category information for question retrieval [7], [19]. On the contrary, we argue that the category information benefits the word embedding learning in this work. The basic idea is that category information encodes the attributes or properties of words, from which we can group similar words according to their categories. Here, a word's category is assigned based on the questions it appeared in. For example, a question "What are the security issues with java?" is under the category of "Computers & Internet ! Security", we simply put the category of a word java as "Computers & Internet !Security".4 Then, we may require the representations of words that belong to the same category to be close to each other.

Let s(wk; wi; ck)) be the similarity score between wk and wi with the category of ck. Under the above assumption, we use the following heuristic to constrain the similar scores:one main category. For example, “Programming & Design” and “Software” are two similar leaf categories under the main category “Computers & Internet”. Questions under the leaf category “Pro-gramming & Design” may also be relevant with questions under the leaf category “Software”. In order to validate this assumption, where is the combination coefficient. Our goal is to maximize the combined objective Jb, which can be optimized using back propagation neural networks. We call this model as metadata powered model (see Fig. 3), and denote it by MB-NET for easy of reference.

3.2.2Enhanced Category Powered Model

The basic category powered model in equations (3) and (4) is based on the same leaf category assumption, with potential rele-vant questions under the similar leaf categories being omitted. As shown in Fig. 2, there exists several similar leaf categories under

In cQA such as Yahoo! Answers, if a word can appear in multiple questions, this word will belong to multiple categories based on the questions it appeared in. Otherwise, the word will belong to a single category.



validate this assumption, we follow the strategy described in [53] and manually check each question from Yahoo! Answers in the labeled data sets mentioned in Section 4.1 shown in Fig. 4, where X axes represents the number of the similar leaf categories that the relevant questions come from, and Y axes represents the proportion of the relevant questions relative to the number of the similar leaf categories. We find that the relevant questions come from the same leaf category only 42%, that is to say, more than half percentage (58%) come from the similar leaf categories. For Baidu data, we have the similar findings as shown in Fig. 5. Based on these observations, we propose an enhanced category powered model by taking into account the relevant questions under the similar leaf categories, and denoted it by ME-NET:

$$S_e(w_k; w_i; c_k) = \frac{1}{A} S_b(w_k; w_i; c_k) + \sum_{c_j \in Related(c_k)} R(c_j ! c_k) S_b(w_k; w_i; c_j)$$

$$c_j \in Related(c_k) \text{ if } R(c_j ! c_k)$$

To estimate the similar probability between two categories, answerer-based and content-based methods used in [54] can be naturally employed. However, we observe that some leaf cate-gories consist of only a small number of questions, which may lead to the data sparseness. In this paper, we propose to leverage topic models for inferring similar probability between two categories.

In cQA (e.g., Yahoo! Answers), when a user asks a new question, the user has to choose a particular category for the question. The cQA system allows the askers to choose only one leaf category for each question. So the most similar category is not allowed to manually assign to each category when adding it in cQA system. This motivates us to automatically calculate the category similarity.

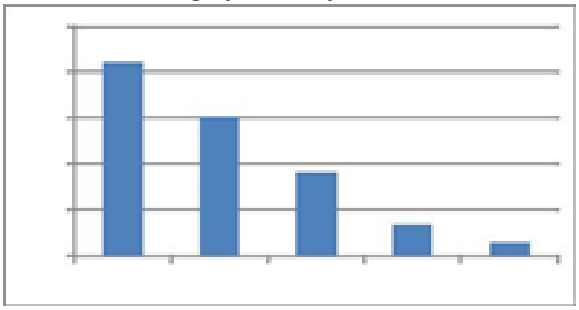
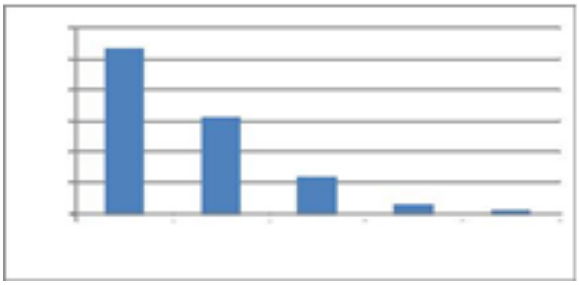


Fig. 4. The proportion of the relevant questions vs. the number of the similar leaf categories from Yahoo! Answers.



3.3Fisher Vector Generation

TABLE 1 Statistics on the manually labeled data.

	#queries	#candidate	#relevant
Yahoo data	1,000	13,000	2,671
Baidu data	1,000	8,000	2,104

of the label is taken as the final decision for a query-candidate pair. We randomly split each of the two labeled data sets into a validation set and a test set with a ration 1 : 3. The validation set is used for tuning parameters of different models, while the test set is used for evaluating how well the models ranked relevant candidates in contrast to irrelevant candidates. Table 1 presents the manually labeled data.

Ranking capability of different methods like the existing work [7], we compare them in a ranking task. This may lose recall for some methods, but it can enable large-scale evaluation.

we employ Mean Average Precision (MAP), Mean Reciprocal Rank (MRR), R-Precision (R-Prec), and Precision at N (P@N) as evaluation measures. These measures are widely used in the literature for question retrieval in cQA [7]:

MAP (Mean Average Precision): For a set of queried questions Q, MAP measures the mean of the average precision for each queried question q for a method M:

$$MAP = \frac{\sum_{q \in Q} AvgP(q)}{Q}$$

$$AvgP(q) = \frac{1}{M_{q,N}} \sum_{j=1}^{M} \frac{N_{M_{q,j}}}{j} 1(M_{q,j})$$

where 1(S) is an indicator function which returns 1 when $M_{q,j}$ is relevant based on the test collection we constructed. $N_{M_{q,j}}$ denotes the number of relevant questions among the top j ranked list returned by M for queried question q, and $M_{q,N}$ denotes the total number of relevant questions of queried question q returned by a method M, and $M_{q,j}$ is the j-th question generated by method M for queried question q. MAP rewards methods that return relevant questions early and also rewards correct ranking of the results.

P@N (Precision@N): For a set of queried questions Q, P @N is the fraction of the top N retrieved questions that are relevant to the queried questions for a method M:

$$P@N = \frac{1}{Q} \sum_{q \in Q} \frac{N_{M_{q,N}}}{N}$$

where $N_{M_{q,N}}$ denotes the number of relevant questions among the top N ranked list returned by a method M for queried question q.

MRR (Mean Reciprocal Rank): MRR is a statistical measure for evaluating any process that produces a list of possible responses to a set of queries, ordered by probability of correctness. The reciprocal rank of a query response is the multiplicative inverse of the rank of first correct answer. The mean reciprocal rank is the average of the reciprocal ranks of results for a set .

Finally, we get the following combined objective Je that incorporates the enhanced category information into the word representation learning process:

$J_e = J(\ ) + E_e$

3.2.3 Optimization Procedure

We optimize the regularization function derived from the metadata of category information along with the training process of the skip-gram model. During the procedure of learning word repre-sentations from the context words in the sliding window, if the central word wk hits the category information, the correspond-ing optimization process of the metadata powered regularization function will be activated. Therefore, we maximize the weighted Euclidean distance between the representation of the central word and that of its similar words according to the objective function in Equations (5) and (11). In our implementation, the optimization is conducted by stochastic gradient descent in a mini-batch mode, whose computational complexity is comparable to that of the optimization procedure of the skip-gram model. Therefore, our metadata powered models (basic category powered model and enhanced category powered model) do not change the convergence of the training process of skip-gram model.

The reciprocal value of the mean reciprocal rank corresponds to the harmonic mean of the ranks.

4.2Comparison Methods

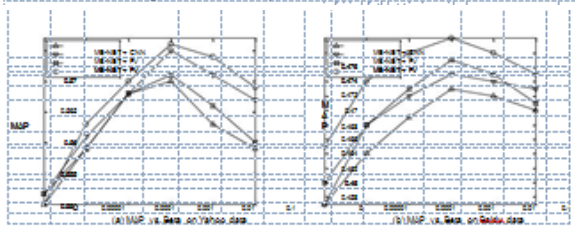


Fig. 6. The performance of MAP for the combination weight on Yahoo data and Baidu data.

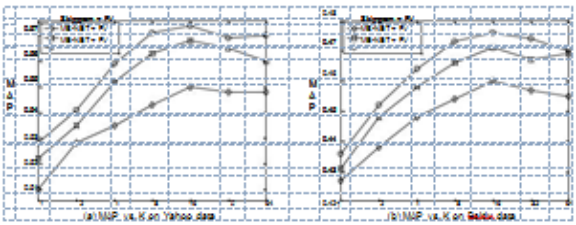


Fig. 7. The performance of MAP for FV with different numbers of Gaus-sians (K) on Yahoo data and Baidu data.

useful information with convolutional and pooling layers, but also learn the matching metrics between the question and its answer.

4.3Parameter Settings

In our experiments, we train the word embeddings on another large-scale data set from cQA sites. For English, we train the word embeddings on the Yahoo! Webscope dataset12. For Chinese, we train the word embeddings on a data set with 1 billion web pages from Baidu Zhidao. These two data sets do not intersect with the above mentioned retrieval data. Little pre-processing is conducted for the training of word embeddings. The resulting text

12. The Yahoo! Webscope dataset Yahoo answers com-prehensive questions and answers version 1.0.2, available at http://reseach.yahoo.com/Academic Relations.

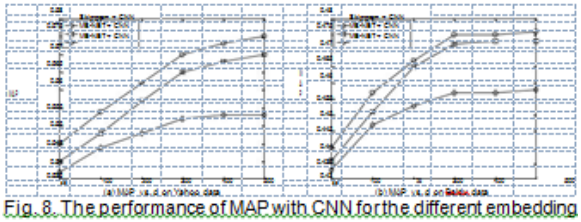


Fig. 8. The performance of MAP with CNN for the different embedding dimension d on Yahoo data and Baidu data.

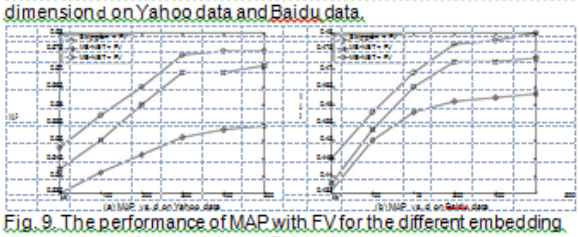


Fig. 9. The performance of MAP with FV for the different embedding dimension d on Yahoo data and Baidu data.

4.4 Experimental Results and Discussions

Now we present the experimental results on the test sets of Yahoo data and Baidu data. In particular, we will compare the baseline word embedding trained by Skip-gram with these methods trained by MB-NET and ME-NET. The dimension of word embedding is set as 300.

Table 2 shows the question retrieval performance in terms of different evaluation metrics. From this table, we can see that learning continuous word embedding representations (MB-NET CNN, ME-NET + CNN, MB-NET + FV, ME-NET + FV) for question retrieval can outperform the syntactic tree matching ap-proach, category-based approaches, translation-based approaches and topic-based approaches on all evaluation metrics. We conduct a statistical test (t-test), the results show that the improvements between the proposed ME-NET + FV and these five groups of compared methods (syntactic tree matching approach, category-based approaches, translation-based approaches, topic-based ap-proaches and deep learning based approaches) are statistically significant ($p < 0:05$). Meanwhile, we also note that DEEP-MATCH and CNTN achieve the comparable performance with the continuous word embedding representations (Skip-gram + CNN, Skip-gram + FV), which indicate the effectiveness of the deep learning methods for question retrieval. Moreover, the metadata of category information powered models (MB-NET + CNN, ME-NET + CNN, MB-NET + FV, ME-NET + FV) outperform the baseline deep learning models (DEEPMATCH and CNTN) and yield the largest improvements. These results can demonstrate that the metadata powered word embedding is of higher quality than the baseline models with no metadata information regularization, and also imply that the word embedding part contributes the most to the performance. Furthermore, we find that ME-NET + CNN converted to lowercase. Since the proposed framework has no limits in using which of the word embedding learning methods, we only consider the following representative methods: Skip-gram (baseline), MB-NET and ME-MET. To train the word embedding using these three methods, we apply the same setting for their common parameters. Specifically, the count of negative samples r is set to 3; the context window size l is set to 5; each model is trained through 1 epoch; the learning rate is initialized as 0.025 and is set to decrease linearly so that it approached zero at the end of training. To train the GMM process of FV, we use the training data mentioned in Section 4.114.

Besides, the combination weight used in MB-NET and ME-NET also plays an important role in producing high quality word embedding. Overemphasizing the weight of the original objective of Skip-gram may result in weakened influence of metadata, while putting too large weight on metadata powered objective may hurt the generality of learned word embedding. Based on our experience, it is a better way to decode the objective combination weight of the Skip-gram model and metadata information based on the scale of their respective

derivatives during optimization. Therefore, we do an experiment on the validation set to determine the best value among f0; 0:00001; ; 0:1g. Fig. 6 shows the evolution of the MAP for the combination weight on Yahoo data and Baidu data. For all these plots, the two group methods are displayed. From the figure, we can see that the best values achieved when setting the combination weight as 0.001.

For parameter K used in FV, we conduct an experiment on the validation data set to determine the optimal value among f1; 2; 4; ; 64g in terms of MAP. Fig. 7 shows the evolution of the MAP for different number of Gaussians (K) on Yahoo data and Baidu data, respectively. For all the these plots, Skip-gram + FV, MB-NET + FV and ME-NET + FV performances are displayed. We test the dimension of the embedding d = 300. All those figures show that the performance of the FV increases up to 16 Gaussians and then reaches a plateau. Therefore, we set the parameter K = 16 in the rest experiments. Then, we conduct an experiment on the validate Data set to determine the optimal value among f50; 100; 200; 300; 400; 500g in terms of MAP. Fig. 9 and Fig. 8 show the evolution of the MAP for different dimension of embedding d on Yahoo data and Baidu data, respectively. From the figure, we can see that the performance increases when setting larger value of d. However, we also observe that the performance increases very slow when d is larger than 300. Therefore, we set d = 300 in order to make the balance between the performance and the computational complexity.

For parameters used in BM25, LM, WTM, WTLM, VSM+!de, LM + QC and VSM + WTLM, we tune these param-eters on the validate data set to determine the best values. For topic-based approaches, we set the symmetric Dirichlet priors as in the literature [62]. We tune the number of topics in f50; 100; 150; 200; 300g and the number of iterations of Gibbs sampling in f100; 200; ; 1000g. The best parameter combina-tions for TMC, UQATM and SQATM are (100; 1000), (50; 1000) and (50; 1000) respectively on the validation sets of Yahoo data and Baidu data. For DEEPMATCH and CNTN, we determine the best values of parameters on our validation set in terms of MAP. In fact, we tune the parameters of all compared methods on the + and ME-NET + FV obtains the better performances than MB-NET + CNN and MB-NET + FV. Surprisingly we find that the comparisons ME-NET + CNN vs. MB-NET + CNN and ME-NET+FV vs. MB-NET + FV are statistically significant with $p < 0:05$ under the evaluation metric of P@1. This is important because a good QA system should return a relevant answer for the top 1 rank. The results indicates that enhanced category information can further help word embedding learning and question retrieval.

For category-based methods, we find that when incorporating the category information into the existing retrieval models (VSM, BM25, LM and WTLM), the retrieval performance can be further improved. The reason may be that category information often encodes the domain knowledge of the words, which can play an important role in term weighting for question .

Translation-based methods significantly outperform VSM, BM25 and LM, which demonstrate that matching questions with the semantically related translation words or phrases from question-answer pairs can effectively address the word lexical gap problem. Besides, we also note that phrase-based translation model is more effective because it captures some contextual information in modeling the translation of phrases as a whole. More precise translation can be determined for phrases than for words. Similar observation has also been found in the previous work [8].

On both data sets, topic-based models achieve comparable per-formance with the translation-based models and but they perform For each queried question, we retrieve the top 100 results from the collection using Lucene with the BM25 scoring function. We consider this ranked list of results as our baseline. We then re-ranked the list using a learning to rank (LTR) framework. LTR aims at automatically creating the ranking model using training data and machine learning techniques. A typical setting is learning to rank is that feature vectors and ranks are given as training data. A ranking model is learned based on the training data and then applied to the Besides, we also note that when we transform the invariable-length vector into a fixed-length vector for hereafter question retrieval, CNN has the slightly worse performance with FV for most evaluation metrics on both data sets except for R-Prec on Baidu data. The reason may be that CNN has so many parameters to train, which causes unstable in robustness on different data sets. Similar

findings have also been observed by [52]. Therefore, we only present the experimental results using FV in section provements over the baseline BM25 score. For example, MRR is increased by 19.28% and 22.86% (row 1 vs. row 5) with bsF V on Yahoo data and Baidu data, respectively. Similar results are shown for R@1. In addition, we also observe that their combinations can further gain improvements (e.g., row 3 vs. row 4; row 5 vs. row 6; row 7 vs. row 8). These results may indicate that the different granularity representations are orthogonal to each other, their combination can better modeling the similarity.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we propose to model and learn continuous vector representations for question retrieval in cQA. In order to model and learn the better word embeddings, we firstly introduce two novel models: one is basic category powered model called MB-NET and the other one is enhanced category powered model called ME-NET, leveraging the category information within cQA pages to learn word representations and alleviate the lexical gap problem. Once the words are embedded into a continuous space, we treat each question as a BoEW. Then, the variable size BoEWs 15.In cQA site, each resolved question has a best answer chosen by the asker from the several candidate answers. Therefore, we can construct the partial-order training data automatically.

unseen test data. Because of the advantages of it offers, LTR has been gained increasing attention in IR. In this paper, we train the LTR model by using the training data of (question, best answer) pairs. The input to our scoring formula is a (query, best answer) pair (q; ba)15. The first type of features we include is the common statistical-based features (sbf) described in [64]. Table 4 summarizes these features, where L1 is the well known BM25 score [11] of the document as calculated by the search engine for a given query. We use it as one of the baseline features, as well as the one of the baseline scorers to compare with in the experiments. The second type is our semantic level similarity feature.

Finally, we experimented with state-of-the-art LTR algorithm SVMRank [65] to determine the final scoring formula. We choose SVMRank [65] as our LTR framework due to its robustness to the training data. For more details, the readers can refer to the related work [65]. For the automatic test sets on Yahoo data and Baidu data, we evaluate the quality of the retrieved results of the various ranking models using MRR and Binary-Recall R@k (the relative number of queried questions with P @K > 0) [66]. Table 5 presents the results of the automatic evaluation.

Looking at Table 5, we first see that statistical-based features (sbf) within a LTR framework improve over the basic Lucene BM25 ranking function. In our test sets, the improvements of incorporating the statistical-based features (sbf) are 5.47% and 8.89% for MRR on Yahoo data and Baidu data, respectively. Furthermore, if instead of statistical-based features, we take the semantic level features as input to LTR, we still gain the im- are aggregated into fixed-length vectors by using FK. Finally, the dot product between FVs are used to calculate the semantic sim-ilarities for question retrieval. Experiments conducted on English and Chinese cQA data sets demonstrate the effectiveness of our approaches. In addition, we further conduct a series of experiments to evaluate our proposed approaches on large-scale data sets. The results demonstrate the robustness of our proposed approaches.

There are some ways in which this work could be further improved. First, since our framework does not rely on which kind of metadata information is used, it can be easily extended to incorporate other metadata information, such as the user ratings, like signals and Poll and Survey signals, into the learning process to obtain more powerful word representations. Second, FK is just one possible way to transform the variable-length vector repre-sentation into a fixed-length vector, a natural avenue for further research would be the use of more powerful algorithms. Third, inspired by the successful application of attention mechanism in NLP community, we plan to model the question retrieval task with a shared attention mechanism.

6.ACKNOWLEDGMENTS

This work was supported by the National Natural Science Founda-tion of China, and also supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada and an NSERC CREATE award. We thank the anonymous

reviewers and associate editor for their insightful comments.

REFERENCES

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

[2] X. Xue, J. Jeon, and W. B. Croft, "Retrieval models for question and answer archives," in Proceedings of the SIGIR, 2008, pp. 475–482.

[3] L. A. Adamic, J. Zhang, E. Bakshy, and M. S. Ackerman, "Knowledge sharing and yahoo answers: Everyone knows something," in Proceedings of the WWW, 2008, pp. 665–674.

[4] H. Duan, Y. Cao, C.-Y. Lin, and Y. Yu, "Searching questions by identifying question topic and question focus," in Proceedings of ACL, 2008, pp. 156–164.

[5] J.-T. Lee, S.-B. Kim, Y.-I. Song, and H.-C. Rim, "Bridging lexical gaps between queries and questions on large online q&a collections with compact translation models," in Proceedings of the EMNLP, 2008, pp. 410–418.

[6] D. Bernhard and I. Gurevych, "Combining lexical semantic resources with question & answer archives for translation-based answer finding," in Proceedings of the ACL, 2009, pp. 728–736.

[7] 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," in Proceedings of the WWW, 2010, pp. 201– 210.

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

[9] A. Singh, "Entity based q&a retrieval," in Proceedings of the EMNLP, 2012, pp. 1266–1277.

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

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

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

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

[14] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Dis-tributed representations of words and phrases and their compositionality," in Proceedings of the NIPS, 2013, pp. 3111–3119.

[15] Z. Lu and H. Li, "A deep architecture for matching short texts," in Proceedings of the NIPS, 2013, pp. 1367–1375.

[16] X. Qiu and X. Huang, "Convolutional neural tensor network architecture for community-based question answering," in Proceedings of the IJCAI, 2015, pp. 1305–1311.

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

[18] Z. Ming, T. Chua, and G. Cong, "Exploring domain-specific term weight in archived question search," in Proceedings of the CIKM, 2010, pp. 1605–1608.

[19] G. Zhou, Y. Chen, D. Zeng, and J. Zhao, "Towards faster and better retrieval models for question search," in Proceedings of the CIKM, 2013,pp. 2139–2148.

[20] S. Clinchant and F. Perronnin, "Aggregating continuous word embed-dings for information retrieval," in Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality, 2013, pp. 100–109.