Learning Continuous Word Embedding with Metadata for Question Retrieval in Community Question Answering

Guangyou Zhou[1], Tingting He[1], Jun Zhao[2], and Po Hu[1]
[1] School of Computer, Central China Normal University, Wuhan 430079, China
[2] National Laboratory of Pattern Recognition, CASIA, Beijing 100190, China
{gyzhou,tthe,phu}@mail.ccnu.edu.cn jzhao@nlpr.ia.ac.cn

Abstract

Community question answering (cQA) has become an important issue due to the popularity of cQA archives on the web. This paper is concerned with the problem of 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 about new challenge for question retrieval in cQA. In this paper, we propose to learn continuous word embeddings with metadata of category information within cQA pages for question retrieval. To deal with the variable size of word embedding vectors, we employ the framework of fisher kernel to aggregated them into the fixed-length vectors. Experimental results on large-scale real world cQA data set show that our approach can significantly outperform state-of-the-art translation models and topic-based models for question retrieval in cQA.

1 Introduction

Over the past few years, a large amount of user-generated content have 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! Answers[1], Live QnA[2], and Baidu Zhidao[3]. 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 (Jeon et al., 2005; Xue et al., 2008) and knowledge mining (Adamic et al., 2008), etc.

One fundamental task for reusing content in cQA is finding 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. Many studies have been done along this line (Jeon et al., 2005; Xue et al., 2008; Duan et al., 2008; Lee et al., 2008; Bernhard and Gurevych, 2009; Cao et al., 2010; Zhou et al., 2011; Singh, 2012; Zhang et al., 2014a). One big challenge for question retrieval in cQA is 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 (Zhang et al., 2014a), 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 (Robertson et al., 1994)) 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 to improve traditional IR models (Jeon et al., 2005; Xue et al., 2008; Zhou et al., 2011). 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)

[1] http://answers.yahoo.com/
[2] http://qna.live.com/
[3] http://zhidao.baidu.com/

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pages 250–259,
Beijing, China, July 26-31, 2015. ©2015 Association for Computational Linguistics

translation probabilities (Jeon et al., 2005; Xue et al., 2008; Zhou et al., 2011). 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 (Zhang et al., 2014a). The second group is the topic-based models (Cai et al., 2011; Ji et al., 2012), 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 representations. Due to the ability of word embeddings, we firstly transform words in a question into continuous vector representations by looking up tables. These word embeddings are learned in advance using a continuous skip-gram model (Mikolov et al., 2013), or other continuous word representation learning methods. 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 (Clinchant and Perronnin, 2013; Sanchez et al., 2013). Through the two steps, the proposed approach 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 test the proposed approach 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 approach can significantly outperform state-of-the-art translation models and topic-based models for question retrieval in cQA.

Our contribution in this paper are three-fold: (1) we represent a question as a bag-of-embedded-words (BoEW) in a continuous space; (2) 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; (3) an empirical verification of the efficacy of the proposed framework on large-scale English and Chinese cQA data.

The rest of this paper is organized as follows. Section 2 summarizes the related work. Section 3 describes our proposed framework for question retrieval. Section 4 reports the experimental results. Finally, we conclude the paper in Section 5.

2 Related Work

2.1 Question Retrieval in cQA

Significant research efforts have been conducted over the years in attempt to improve question retrieval in cQA (Jeon et al., 2005; Xue et al., 2008; Lee et al., 2008; Duan et al., 2008; Bernhard and Gurevych, 2009; Cao et al., 2010; Zhou et al., 2011; Singh, 2012; Zhang et al., 2014a). Most of these works focus on finding similar questions for the user queried questions. The major challenge for question retrieval in cQA is the lexical gap problem. Jeon et al. (2005) proposed a word-based translation model for automatically fixing the lexical gap problem. Xue et al. (2008) proposed a word-based translation language model for question retrieval. Lee et al. (2008) tried to further improve the translation probabilities based on question-answer pairs by selecting the most important terms to build compact translation models. Bernhard and Gurevych (2009) proposed to use as a parallel training data set the definitions and glosses provided for the same term by different lexical semantic resources. In order to improve the word-based translation model with some contextual information, Riezler et al. (2007) and Zhou et al. (2011) 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. Singh (2012) addressed the lexical gap issues by extending the lexical word-based translation model to incorporate semantic information (entities).

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 questions and answers with deep linguistic analysis (Duan et al., 2008; Wang et al., 2009; Wang et al., 2010; Ji et al., 2012; Zhang et al., 2014a) or a learning to rank strategy (Surdeanu et al., 2008; Carmel et al., 2014). Recently, Cao et al. (2010) and Zhou et al. (2013) exploited the category metadata within cQA pages to further improve the performance. On the contrary, we focus on the representation learning for questions, with a different solution with those previous works.

2.2 Word Embedding Learning

Representation of words as continuous vectors has attracted increasing attention in the area of natural language processing (NLP). Recently, a series of works applied deep learning techniques to learn high-quality word representations. Bengio et al. (2003) proposed a probabilistic neural network language model (NNLM) for word representations. Furthermore, Mikolov et al. (2013) proposed efficient neural network models for learning 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 (Huang et al., 2012; Maas et al., 2011; Turian et al., 2010).

Nevertheless, since most the existing works learned word representations 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. (2014) used semantic prior knowledge to improve word representations. Xu et al. (2014) 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 between 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 estimation 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

In this paper, we consider a context-aware predicting model, more specifically, the *Skip-gram* model (Mikolov et al., 2013) for learning word embeddings, since it is much more efficient as well as memory-saving than other approaches.[4] *Skip-gram* is recently proposed for learning word representations using a neural network model, whose underlying idea is that similar words should have similar contexts. In the *Skip-gram* model (see Figure 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 w_k as input to predict the context words. Let $M_{d \times N}$ denote the learned embedding matrix,

[4] Note that although we use the skip-gram model as an example to illustrate our approach, the similar framework can be developed on the basis of any other word embedding models.

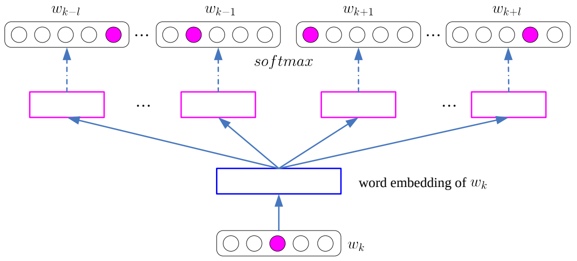

Figure 1: The continuous skip-gram model.

where N is the vocabulary size and d is the dimension of word embeddings. Each column of M represents the embedding of a word. Let w_k is first mapped to its embedding e_{w_k} by selecting the corresponding column vector of M. The probability of its context word w_{k+j} is then computed using a log-linear *softmax* function:

$$p(w_{k+j}|w_k;\theta) = \frac{\exp(e^T_{w_{k+j}} e_{w_k})}{\sum_{w=1}^{N} \exp(e^T_{w} e_{w_k})} \qquad (1)$$

where θ are the parameters we should learned, $k = 1 \cdots d$, and $j \in [-l, l]$. Then, the log-likelihood over the entire training data can be computed as:

$$J(\theta) = \sum_{(w_k, w_{k+j})} \log p(w_{k+j}|w_k;\theta) \qquad (2)$$

To calculate the prediction errors for back propagation, we need to compute the derivative of $p(w_{k+j}|w_k;\theta)$, whose computation 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. (2013) 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 $p(w) \propto u(w)^{3/4}$ (Mikolov et al., 2013).

3.2 Metadata Powered Model

After briefing the *skip-gram* model, we introduce how we equip it with the metadata information. In cQA sites, there are several metadata, such as "category","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. When an user asks a question, the user typically required to choose a category label for the question from a predefined hierarchy of categories (Cao et al., 2010; Zhou et al., 2013). Previous work in the literature has demonstrated the effectiveness of the category information for question retrieval (Cao et al., 2010; Zhou et al., 2013). 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". Then, we may require the representations of words that belong to the same category to be close to each other.

Let $s(w_k, w_i)$ be the similarity score between w_k and w_i. Under the above assumption, we use the following heuristic to constrain the similar scores:

$$s(w_k, w_i) = \begin{cases} 1 & \text{if } c(w_k) = c(w_i) \\ 0 & \text{otherwise} \end{cases} \qquad (3)$$

where $c(w_k)$ denotes the category of w_k. If the central word w_k shares the same category with the word w_i, their similarity score will become 1, otherwise, we set to 0. Then we encode the category information using a regularization function E_c:

$$E_c = \sum_{k=1}^{N} \sum_{i=1}^{N} s(w_k, w_i) d(w_k, w_i) \qquad (4)$$

where $d(w_k, w_i)$ is the distance for the words in the embedding space and $s(w_k, w_i)$ serves as a weighting function. Again, for simplicity, we define $d(w_k, w_i)$ as the Euclidean distance between w_k and w_i.

We combine the *skip-gram* objective function and the regularization function derived from the metadata of category information, we get the following combined objective J_c that incorporates category information into the word representation learning process:

$$J_c = J(\theta) + \beta E_c \qquad (5)$$

where β is the combination coefficient. Our goal is to maximize the combined objective J_c, which

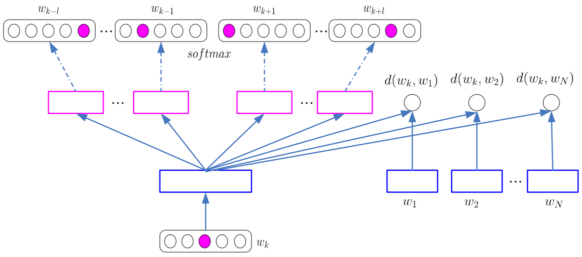

Figure 2: The continuous skip-gram model with metadata of category information, called M-NET.

can be optimized using back propagation neural networks. We call this model as metadata powered model (see Figure 2), and denote it by M-NET for easy of reference.

In the implementation, 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 representations from the context words in the sliding window, if the central word w_k hits the category information, the corresponding 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 Equation (5).

3.3 Fisher Vector Generation

Once the word embeddings are learned, questions can be represented by variable length sets of word embedding vectors, which can be viewed as BoEWs. Semantic level similarities between queried questions and the existing questions represented by BoEWs can be captured more accurately than previous bag-of-words (BoW) methods. However, since BoEWs are variable-size sets of word embeddings and most of the index methods in information retrieval field are not suitable for this kinds of issues, BoEWs cannot be directly used for large-scale question retrieval task.

Given a cQA data collection $\mathcal{Q} = \{q_i, 1 \leq i \leq |\mathcal{Q}|\}$, where q_i is the ith question and $|\mathcal{Q}|$ is the number of questions in the data collection. The ith question q_i is composed by a sequence of words $w_i = \{w_{ij}, 1 \leq j \leq N_i\}$, where N_i denotes the length of q_i. Through looking up table (word embedding matrix) of M, the ith question q_i can be represented by $E_{w_i} = \{e_{w_{ij}}, 1 \leq j \leq N_i\}$, where $e_{w_{ij}}$ is the word embedding of w_{ij}. According to the framework of FK (Clinchant and Perronnin, 2013; Sanchez et al., 2013; Zhang et al., 2014b), questions are modeled by a probability density function. In this work, we use Gaussian mixture model (GMM) to do it. We assume that the continuous word embedding E_{w_i} for question q_i have been generated by a "universal" (e.g., question-independent) probability density function (pdf). As is a common practice, we choose this pdf to be a GMM since any continuous distribution can be approximated with arbitrary precision by a mixture of Gaussian. In what follows, the pdf is denoted u_λ where $\lambda = \{\theta_i, \mu_i, \Sigma_i, i = 1 \cdots K\}$ is the set of parameters of the GMM. θ_i, μ_i and Σ_i denote respectively the mixture weight, mean vector and covariance matrix of Gaussian i. For computational reasons, we assume that the covariance matrices are diagonal and denote σ_i^2 the variance vector of Gaussian i, e.g., $\sigma_i^2 = \text{diag}(\sum_i)$. In real applications, the GMM is estimated offline with a set of continuous word embeddings extracted from a representative set of questions. The parameters λ are estimated through the optimization of a Maximum Likelihood (ML) criterion using the Expectation-Maximization (EM) algorithm. In the following, we follow the notations used in (Sanchez et al., 2013).

Given u_λ, one can characterize the question q_i using the following score function:

$$G_\lambda^{q_i} = \bigtriangledown_\lambda^{N_i} \text{log} u_\lambda(q_i) \tag{6}$$

where $G_\lambda^{q_i}$ is a vector whose size depends only on the number of parameters in λ. Assuming that the word embedding $e_{w_{ij}}$ is iid (a simplifying assumption), we get:

$$G_\lambda^{q_i} = \sum_{j=1}^{N_i} \bigtriangledown_\lambda \text{log} u_\lambda(e_{w_{ij}}) \tag{7}$$

Following the literature (Sanchez et al., 2013), we propose to measure the similarity between two questions q_i and q_j using the FK:

$$K(q_i, q_j) = G_\lambda^{q_i^T} F_\lambda^{-1} G_\lambda^{q_j} \tag{8}$$

where F_λ is the Fisher Information Matrix (FIM) of u_λ:

$$F_\lambda = E_{q_i \sim u_\lambda} \left[G_\lambda^{q_i} G_\lambda^{q_i^T} \right] \tag{9}$$

Since F_λ is symmetric and positive definite, F_λ^{-1} can be transformed to $L_\lambda^T L_\lambda$ based on the Cholesky decomposition. Hence, $K_{FK}(q_i, q_j)$ can rewritten as follows:

$$K_{FK}(q_i, q_j) = \mathcal{G}_\lambda^{q_i^T} \mathcal{G}_\lambda^{q_j} \tag{10}$$

where

$$\mathcal{G}_{\lambda}^{q_i} = L_{\lambda} G_{\lambda}^{q_i} = L_{\lambda} \bigtriangledown_{\lambda} \log u_{\lambda}(q_i) \tag{11}$$

In (Sanchez et al., 2013), $\mathcal{G}_{\lambda}^{q_i}$ refers to as the *Fisher Vector* (FV) of q_i. The dot product between FVs can be used to calculate the semantic similarities. Based on the specific probability density function, GMM, FV of q_i is respect to the mean μ and standard deviation σ of all the mixed Gaussian distributions. Let $\gamma_j(k)$ be the soft assignment of the jth word embedding $e_{w_{ij}}$ in q_i to Guassian k (u_k):

$$\gamma_j(k) = p(k|e_{w_{ij}}) \frac{\theta_i u_k(e_{w_{ij}})}{\sum_{j=1}^{K} \theta_k u_k(e_{w_{ij}})} \tag{12}$$

Mathematical derivations lead to:

$$\mathcal{G}_{\mu,k}^{q_i} = \frac{1}{N_i \sqrt{\theta_i}} \sum_{j=1}^{N_i} \gamma_j(k) \Big[\frac{e_{w_{ij}} - \mu_k}{\sigma_k} \Big] \tag{13}$$

$$\mathcal{G}_{\sigma,k}^{q_i} = \frac{1}{N_i \sqrt{2\theta_i}} \sum_{j=1}^{N_i} \gamma_j(k) \Big[\frac{(e_{w_{ij}} - \mu_k)^2}{\sigma_k^2} - 1 \Big]$$

The division by the vector σ_k should be understood as a term-by-term operation. The final gradient vector $\mathcal{G}_{\lambda}^{q_i}$is the concatenation of the $\mathcal{G}_{\mu,k}^{q_i}$ and $\mathcal{G}_{\sigma,k}^{q_i}$ vectors for $k = 1 \cdots K$. Let d denote the dimensionality of the continuous word embeddings and K be the number of Gaussians. The final fisher vector $\mathcal{G}_{\lambda}^{q_i}$ is therefore $2Kd$-dimensional.

4 Experiments

In this section, we present the experiments to evaluate the performance of the proposed method for question retrieval.

4.1 Data Set and Evaluation Metrics

We collect the data sets from Yahoo! Answers and Baidu Zhidao. Yahoo! Answers and Baidu Zhidao represent the largest and the most popular cQA archives in English and Chinese, respectively. More specifically, we utilized the *resolved* questions at Yahoo! Answers and Baidu Zhidao. The questions include 10 million items from Yahoo! Answers and 8 million items from Baidu Zhidao (also called retrieval data). Each resolved question consists of three fields: "title", "description" and "answers", as well as some metadata, such as "category". For question retrieval, we use only the "title" field and "category" metadata. It is assumed that the titles of questions already provide enough semantic information for understanding users' information needs (Duan et al., 2008). We develop two test sets, one for "Yahoo data", and the other for "Baidu data". In order to create the test sets, we collect some extra questions that have been posted more recently than the retrieval data, and randomly sample $1,000$ questions for Yahoo! Answers and Baidu Zhidao, respectively. We take those questions as queries. All questions are lowercased and stemmed. Stopwords[5] are also removed.

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

Table 1: Statistics on the manually labeled data.

We separately index all data from Yahoo! Answers and Baidu Zhidao using an open source *Lucene* with the BM25 scoring function[6]. For each query from Yahoo! Answers and Baidu Zhidao, we retrieve the several candidate questions from the corresponding indexed data by using the BM25 ranking algorithm in *Lucene*. On average, each query from Yahoo! Answers has 13 candidate questions and the average number of candidate questions for Baidu Zhidao is 8.

We recruit students to label the relevance of the candidate questions regarding to the queries. Specifically, for each type of language, we let three native students. Given a candidate question, a student is asked to label it with "relevant" or "irrelevant". If a candidate question is considered semantically similar to the query, the student will label it as "relevant"; otherwise, the student will label it as "irrelevant". As a result, each candidate question gets three labels and the majority 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.

Please note that rather than evaluate both retrieval and ranking capability of different meth-

[5] http://truereader.com/manuals/onix/stopwords1.html

[6] We use the BM25 implementation provided by Apache Lucene (http://lucene.apache.org/), using the default parameter setting ($k_1 = 1.2$, $b = 0.75$)

ods like the existing work (Cao et al., 2010), we compare them in a ranking task. This may lose recall for some methods, but it can enable large-scale evaluation.

In order to evaluate the performance of different models, we employ Mean Average Precision (MAP), Mean Reciprocal Rank (MRR), R-Precision (R-Prec), and Precision at K (P@5) as evaluation measures. These measures are widely used in the literature for question retrieval in cQA (Cao et al., 2010).

4.2 Parameter Setting

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 dataset[7]. 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 is tokenized using the Stanford tokenizer,[8], and every word is converted to lowercase. Since the proposed framework has no limits in using which of the word embedding learning methods, we only consider the following two representative methods: *Skip-gram* (baseline) and *M-NET*. To train the word embedding using these two 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.

Besides, the combination weight β used in *M-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. Finally, we set $\beta = 0.001$ empirically. Note that if the parameter is optimized on the validation set, the final performance can be further improved.

For parameter K used in FV, we do an experiment on the validation data set to determine the best value among 1, 2, 4, $\cdots$, 64 in terms of MAP. As a result, we set $K = 16$ in the experiments empirically as this setting yields the best performance.

4.3 Main Results

In this subsection, we present the experimental results on the test sets of Yahoo data and Baidu data. We compare the baseline word embedding trained by *Skip-gram* against this trained by *M-NET*. The dimension of word embedding is set as 50,100 and 300. Since the motivation of this paper attempts to tackle the lexical gap problem for queried questions and questions in the archive, we also compare them with the two groups of methods which also address the lexical gap in the literature. The first group is the translation models: word-based translation model (Jeon et al., 2005), word-based translation language model (Xue et al., 2008), and phrase-based translation model (Zhou et al., 2011). We implement those three translation models based on the original papers and train those models with (question, best answer) pairs from the Yahoo! Webscope dataset Yahoo answers and the 1 billion web pages of Baidu Zhidao for English and Chinese, respectively. Training the translation models with different pairs (e.g., question-best answer, question-description, question-answer) may achieve inconsistent performance on Yahoo data and Baidu data, but its comparison and analysis are beyond the scope of this paper. The second group is the topic-based methods: unsupervised question-answer topic model (Ji et al., 2012) and supervised question-answer topic model (Zhang et al., 2014a). We re-implement these two topic-based models and tune the parameter settings on our data set. Besides, we also introduce a baseline language model (LM) (Zhai and Lafferty, 2001) for comparison.

Table 2 shows the question retrieval performance by using different evaluation metrics. From this table, we can see that learning continuous word embedding representations (Skip-gram + FV, M-NET + FV) for question retrieval can outperform the translation-based approaches and topic-based approaches on all evaluation metrics. We conduct a statistical test (t-test), the results

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

[8] http://nlp.stanford.edu/software/tokenizer.shtml

Model	dim	Yahoo data				Baidu data			
		MAP	MRR	R-Prec	P@5	MAP	MRR	R-Prec	P@5
LM (baseline)	-	0.435	0.472	0.381	0.305	0.392	0.413	0.325	0.247
(Jeon et al., 2005)	-	0.463	0.495	0.396	0.332	0.414	0.428	0.341	0.256
(Xue et al., 2008)	-	0.518	0.560	0.423	0.346	0.431	0.435	0.352	0.264
(Zhou et al., 2011)	-	0.536	0.587	0.439	0.361	0.448	0.450	0.367	0.273
(Ji et al., 2012)	-	0.508	0.544	0.405	0.324	0.425	0.431	0.349	0.258
(Zhang et al., 2014a)	-	0.527	0.572	0.433	0.350	0.443	0.446	0.358	0.265
Skip-gram + FV	50	0.532	0.583	0.437	0.358	0.447	0.450	0.366	0.272
	100	0.544	0.605†	0.440	0.363	0.454	0.457	0.373	0.274
	300	0.550†	0.619†	0.444	0.365	0.460†	0.464†	0.374	0.277
M-NET + FV	50	0.548†	0.612†	0.441	0.363	0.459†	0.462†	0.374	0.276
	100	0.562‡	0.628‡	0.452†	0.367‡	0.468‡	0.471	0.378†	0.280†
	300	**0.571**‡	**0.643**‡	**0.455**‡	**0.374**‡	**0.475**‡	**0.477**‡	**0.385**‡	**0.283**‡

Table 2: Evaluation results on Yahoo data and Baidu data, where dim denotes the dimension of the word embeddings. The bold formate indicates the best results for question retrieval. † indicates that the difference between the results of our proposed approach (Skip-gram + FV, M-NET + FV) and other methods are mildly significant with $p < 0.08$ under a t-test; ‡ indicates the comparisons are statistically significant with $p < 0.05$.

show that the improvements between the proposed M-NET + FV and the two groups of compared methods (translation-based approaches and topic-based approaches) are statistically significant ($p < 0.05$), while the improvements between Skip-gram + FV and the translation-based approaches are mildly significant ($p < 0.08$). Moreover, the metadata of category information powered model (M-NET + FV) outperforms the baseline skip-gram model (Skip-gram + FV) and yields the largest improvements. These results can imply that the metadata powered word embedding is of higher quality than the baseline model with no metadata information regularization. Besides, we also note that setting higher dimension brings more improvements for question retrieval task.

Translation-based methods significantly outperform 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 (Zhou et al., 2011).

On both data sets, topic-based models achieve comparable performance with the translation-based models and but they perform better than LM. The results demonstrate that learning the latent topics aligned across the question-answer pairs can be an alternative for bridging lexical gap problem for question retrieval.

5 Conclusion

This paper proposes to learn continuous vector representations for question retrieval in cQA. We firstly introduce a new metadata powered word embedding method, called M-NET, to leverage the category information within cQA pages to obtain word representations. Once the words are embeded in a continuous space, we treat each question as a BoEW. Then, the variable size BoEWs are aggregated into fixed-length vectors by using FK. Finally, the dot product between FVs are used to calculate the semantic similarities for question retrieval. Experiments on large-scale real world cQA data demonstrate that the efficacy of the proposed approach. For the future work, we will explore how to incorporate more types of metadata information, such as the *user ratings*, *like signals* and *Poll and Survey signals*, into the learning process to obtain more powerful word representations.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 61303180, No. 61272332 and 61402191), the Beijing Natural Science Foundation (No. 4144087), the Major Project of National Social Science Found (No. 12&2D223), the Fundamental Research Funds for the Central Universities (No. CCNU15ZD003), and also Sponsored by CCF-Tencent Open Research Fund. We thank the anonymous reviewers for their insightful comments.

References

Lada A. Adamic, Jun Zhang, Eytan Bakshy, and Mark S. Ackerman. 2008. Knowledge sharing and yahoo answers: Everyone knows something. In *Proceedings of WWW*, pages 665–674.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3.

Delphine Bernhard and Iryna Gurevych. 2009. Combining lexical semantic resources with question & answer archives for translation-based answer finding. In *Proceedings of ACL-IJCNLP*.

Li Cai, Guangyou Zhou, Kang Liu, and Jun Zhao. 2011. Learning the latent topics for question retrieval in community qa. In *Proceedings of IJCNLP*, pages 273–281.

Xin Cao, Gao Cong, Bin Cui, and Christian S. Jensen. 2010. A generalized framework of exploring category information for question retrieval in community question answer archives. In *Proceedings of WWW*, pages 201–210.

David Carmel, Avihai Mejer, Yuval Pinter, and Idan Szpektor. 2014. Improving term weighting for community question answering search using syntactic analysis. In *Proceedings of CIKM*, pages 351–360.

Stephane Clinchant and Florent Perronnin. 2013. Aggregating continuous word embeddings for information retrieval. In *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*, pages 100–109.

Huizhong Duan, Yunbo Cao, Chin yew Lin, and Yong Yu. 2008. Searching questions by identifying question topic and question focus. In *Proceedings of ACL*.

Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of ACL*, pages 873–882.

Jiwoon Jeon, W. Bruce Croft, and Joon Ho Lee. 2005. Finding similar questions in large question and answer archives. In *Proceedings of CIKM*.

Zongcheng Ji, Fei Xu, Bin Wang, and Ben He. 2012. Question-answer topic model for question retrieval in community question answering. In *Proceedings of CIKM*, pages 2471–2474.

Jung-Tae Lee, Sang-Bum Kim, Young-In Song, and Hae-Chang Rim. 2008. Bridging lexical gaps between queries and questions on large online q&a collections with compact translation models. In *Proceedings of EMNLP*, pages 410–418.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of ACL*, pages 142–150.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*, pages 3111–3119.

Stefan Riezler, Er Vasserman, Ioannis Tsochantaridis, Vibhu Mittal, and Yi Liu. 2007. Statistical machine translation for query expansion in answer retrieval. In *Proceedings of ACL*.

S. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. 1994. Okapi at trec-3. In *Proceedings of TREC*, pages 109–126.

Jorge Sanchez, Florent Perronnin, Thomas Mensink, and Jakob J. Verbeek. 2013. Image classification with the fisher vector: Theory and practice. *International Journal of Computer Vision*, pages 222–245.

A. Singh. 2012. Entity based q&a retrieval. In *Proceedings of EMNLP*, pages 1266–1277.

M. Surdeanu, M. Ciaramita, and H. Zaragoza. 2008. Learning to rank answers on large online qa collections. In *Proceedings of ACL*, pages 719–727.

Joseph Turian, Lev-Arie Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of ACL*.

Kai Wang, Zhaoyan Ming, and Tat-Seng Chua. 2009. A syntactic tree matching approach to finding similar questions in community-based qa services. In *Proceedings of SIGIR*, pages 187–194.

B. Wang, X. Wang, C. Sun, B. Liu, and L. Sun. 2010. Modeling semantic relevance for question-answer pairs in web social communities. In *ACL*.

Chang Xu, Yalong Bai, Jiang Bian, Bin Gao, Gang Wang, Xiaoguang Liu, and Tie-Yan Liu. 2014. Rc-net: A general framework for incorporating knowledge into word representations. In *Proceedings of CIKM*, pages 1219–1228.

Xiaobing Xue, Jiwoon Jeon, and W. Bruce Croft. 2008. Retrieval models for question and answer archives. In *Proceedings of SIGIR*, pages 475–482.

Mo Yu and Mark Dredze. 2014. Improving lexical embeddings with semantic knowledge. In *Proceedings of ACL*, pages 545–550.

Chengxiang Zhai and John Lafferty. 2001. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of SIGIR*, pages 334–342.

Kai Zhang, Wei Wu, Haocheng Wu, Zhoujun Li, and Ming Zhou. 2014a. Question retrieval with high quality answers in community question answering. In *Proceedings of CIKM*, pages 371–380.

Qi Zhang, Jihua Kang, Jin Qian, and Xuanjing Huang. 2014b. Continuous word embeddings for detecting local text reuses at the semantic level. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 797–806.

Guangyou Zhou, Li Cai, Jun Zhao, and Kang Liu. 2011. Phrase-based translation model for question retrieval in community question answer archives. In *Proceedings of ACL*, pages 653–662.

Guangyou Zhou, Yubo Chen, Daojian Zeng, and Jun Zhao. 2013. Towards faster and better retrieval models for question search. In *Proceedings of CIKM*, pages 2139–2148.