Learning the Latent Topics for Question Retrieval in Community QA

Li Cai, Guangyou Zhou, Kang Liu, and Jun Zhao
National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences
95 Zhongguancun East Road, Beijing 100190, China
{lcai,gyzhou,kliu,jzhao}@nlpr.ia.ac.cn

Abstract

Community-based Question Answering (cQA) is a popular online service where users can ask and answer questions on any topics. This paper is concerned with the problem of question retrieval. Question retrieval in cQA aims to find historical questions that are semantically equivalent or relevant to the queried questions. Although the translation-based language model (Xue et al., 2008) has gained the state-of-the-art performance for question retrieval, they ignore the latent topic information in calculating the semantic similarity between questions. In this paper, we propose a topic model incorporated with the category information into the process of discovering the latent topics in the content of questions. Then we combine the semantic similarity based latent topics with the translation-based language model into a unified framework for question retrieval. Experiments are carried out on a real world cQA data set from Yahoo! Answers. The results show that our proposed method can significantly improve the question retrieval performance of translation-based language model.

1 Introduction

Over the past few years, large scale question and answer archives have become an important information resource on the Web. These include the traditional FAQ archives constructed by the experts or companies for their products and the emerging community-based online services, such as Yahoo! Answers[1] and Live QnA[2].

The major challenge for cQA retrieval is the lexical gap (or *lexical chasm*) between the queried questions and the question-answer pairs in the archives (Jeon et al., 2005; Xue et al., 2008). To solve the *lexical gap* problem, most researchers regarded the question retrieval task as a statistical machine translation problem by using IBM model 1 (Brown et al., 1993) to learn the word-to-word translation probabilities (Berger and Lafferty, 1999; Jeon et al., 2005; Xue et al., 2008; Lee et al., 2008; Bernhard and Gurevych, 2009; Cao et al., 2010). Although the translation-based language model(TRLM) has yielded the state-of-the-art performance for question retrieval, they model the word translation probabilities without taking into account the distribution of words in the whole content.

In this paper, we argue that it is beneficial to exploit the latent topic information for question retrieval. The basic idea is as follows: first we employ the topic model (e.g., LDA) to discover the latent topics in the content of questions, and calculate the semantic similarity between questions based on the latent topic information. Moreover, a distinctive feature of question-answer archives in cQA is that cQA services always organize questions into a hierarchy of categories. We propose an improved latent topic model by introducing the category information of questions. To solve the *lexical gap* problem, the translation-based language model extracts knowledge from question-answer pairs which are collected from cQA service. Latent topic model extracts knowledge from the distribution of words and categories in whole cQA archives. We assume that the two knowledge are complementary to each other, as we will show in the experiment.

In order to illustrate the above ideas clearly, we give an example of retrieving semantically equivalent or relevant to the queried questions in Figure 1. Given question $Q1$, we get a ranked list of semantically similar questions (Q_2, Q_3, Q_4, Q_5) using state-of-the-art translation-based lan-

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

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 273–281,
Chiang Mai, Thailand, November 8 – 13, 2011.

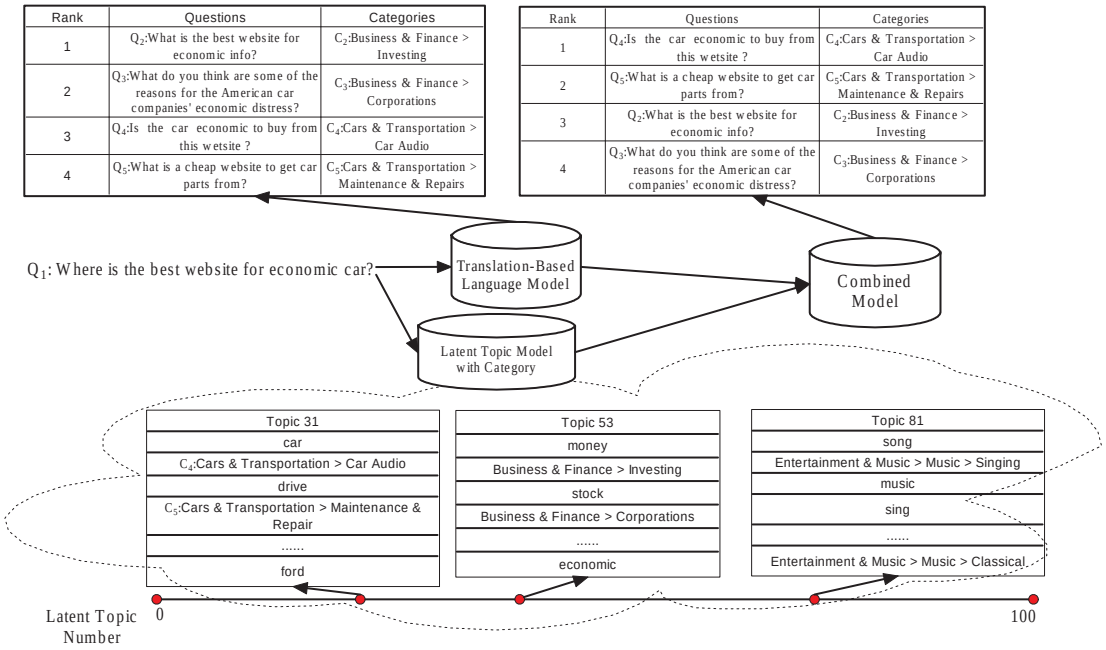

Figure 1: Illustration of our proposed approach.

guage model. All the semantically similar questions are with their corresponding categories. Our proposed latent topic model models the distribution of words, categories of the whole content. We illustrate in Figure 1 a matching of the top words and categories from a few topics. We can see that the word *car* is more related to categories "Cars & Transportation>Car Audio" and "Cars & Transportation>Maintenance & Repairs" than the word "economic" to categories "Business & Finance>Investing" and "Business & Finance>Corporations" in latent topics. Using this information from the latent topic model, we can rerank the retrieved question. Therefore, combining the translation-based language model with the latent topic model with categories, we can get the ranked list of semantically similar questions (Q_4, Q_5, Q_2, Q_3) which are better than the previous retrieval result.

Specifically, our contributions are as follows:

1. We employ the topic model to discover the latent topic information in the content of questions for cQA retrieval (in Section 4.1.)

2. We introduce the category information into the process of discovering the latent topics. (in Section and 4.2).

3. We propose to combine the semantic similarity based latent topics with the translation-based language model into a unified framework to further improve the retrieval performance (in Section 4.4).

4. Finally, we conduct the experiments on cQA data set from Yahoo! Answers for question retrieval. The results show that our proposed approach significantly outperform the state-of-the-art translation-based language model (in Section 5).

The remainder of this paper is organized as follows. Section 2 reviews the related work on community-based question retrieval. Section 3 presents the existing question retrieval models. Section 4 presents the topic model incorporated with category information for question retrieval. Section 5 presents the experimental results. Finally, we conclude and offer the further work in Section 6.

2 Related Work

Recently, the research of question retrieval has been further extended to the cQA data. Jeon et al. (2005) proposed a word-based translation model for automatically fixing the lexical gap problem. Experimental results demonstrated that translation model significantly outperformed the traditional methods (i.e., VSM, BM25, LM). Xue et al. (2008) proposed a translation-based language model for question retrieval. The results indicated that translation-based language model further improved the retrieval results and obtained the

state-of-the-art performance.

Subsequent work on translation models focused on providing suitable parallel data to learn the translation probabilities. 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. Cao et al. (2010) explored adding the category information into the translation model for question retrieval. Zhou et al. (2011) proposed a phrase-based translation model for question retrieval and obtained the state-of-the-art performance.

However, all the existing methods ignore the latent topics information in calculating the semantic similarity between questions. In this paper, we present a new approach to discover the latent topic of questions for improving the performance of translation-based language models for question retrieval. Moreover, we introduce the category information into the process of discovering the latent topics. To the best of our knowledge, none of the existing studies addressed question retrieval in cQA by learning the latent topics.

3 Preliminaries

3.1 Language Model

The unigram language model has been widely used for question retrieval on community-based Q&A data (Jeon et al., 2005; Xue et al., 2008; Cao et al., 2010). To avoid zero probability, we use Jelinek-Mercer smoothing (Zhai and Lafferty, 2001) due to its good performance and cheap computational cost. So the ranking function for the query likelihood language model with Jelinek-Mercer smoothing can be written as:

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

$$P_{ml}(w|Q) = \frac{\#(w,Q)}{|Q|}, \;\; P_{ml}(w|C) = \frac{\#(w,C)}{|C|} \quad (2)$$

where $\mathbf{q}$ is the queried question, Q is a historical question, C is background collection, λ is smoothing parameter. $\#(t,Q)$ is the frequency of term t in Q, $|Q|$ and $|C|$ denote the length of Q and C, respectively.

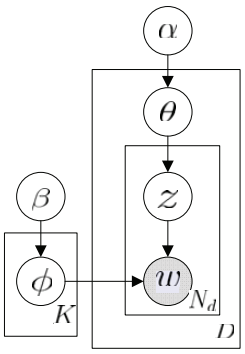

Figure 2: Latent Dirichlet Allocation.

3.2 Translation Model

Previous work (Berger et al., 2000; Jeon et al., 2005; Xue et al., 2008) consistently reported that the word-based translation models (TR) yielded better performance than the traditional methods (VSM, Okapi and LM) for question retrieval. These models exploited the word translation probabilities in a language modeling framework. According to Jeon et al. (2005) and Xue et al. (2008), the ranking function can be written as:

$$P_{TR}(q|Q) = \prod_{w \in \mathbf{q}} (1-\lambda) P_{tr}(w|Q) + \lambda P_{ml}(w|C) \quad (3)$$

$$P_{tr}(w|Q) = \sum_{t \in Q} P(w|t) P_{ml}(t|Q), \;\; P_{ml}(t|Q) = \frac{\#(t,Q)}{|Q|} \quad (4)$$

where $P(w|t)$ denotes the translation probability from word t to word w.

3.3 Translation-Based Language Model

Xue et al. (2008) proposed to linearly mix two different estimations by combining language model and translation model into a unified framework, called TRLM. The experiments show that this model gains better performance than both the language model and the translation model. Following Xue et al. (2008), this model can be written as:

$$P_{TRLM}(q|Q) = \prod_{w \in \mathbf{q}} (1-\lambda) P_{mx}(w|Q) + \lambda P_{ml}(w|C) \quad (5)$$

$$P_{mx}(w|Q) = \delta \sum_{t \in Q} P(w|t) P_{ml}(t|Q) + (1-\delta) P_{ml}(w|Q) \quad (6)$$

4 Topic Model Incorporated with Category Information for Question Retrieval

Previous work on question retrieval in cQA, employs different retrieval models, such as

Symbol	Description
K	the number of topics
N	the number of questions
$\|V\|$	the number of unique words
$\|C\|$	the number of unique leaf categories
N_q	the number of distinct words in question q
θ_q	multinomial distribution over topics specific to question q
ϕ_z	multinomial distribution over words specific to topic z
ψ_z	multinomial distribution over categories specific to topic z
z_{qi}	the topic of the ith word in question q
c_{qi}	the category of the ith word in question q
w_{qi}	the ith word in question q

Table 1: Meanings of the notations used in this paper

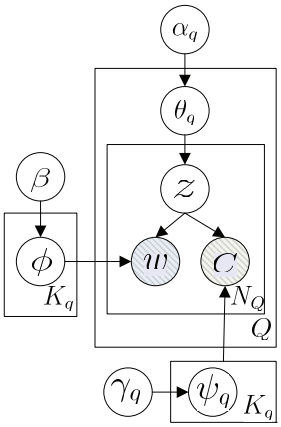

Figure 3: Topic model incorporated with category information.

VSM (Salton et al., 1975), LM (Zhai and Lafferty, 2001), TR (Jeon et al., 2005) and TRLM (Xue et al., 2008). However, all these existing models ignore the latent topics in calculating the semantic similarity between questions. In this Section, we explore the latent topic information for question retrieval.

4.1 Topic Model for Question Retrieval

Before introducing our proposed method, we first briefly describe the basic Latent Dirichlet Allocation (LDA) model (Blei et al., 2003). The notations we used in this paper are presented in Table 1, and the graphic model representations of LDA model is shown in Figure 2. LDA models the generation of document content as two independent stochastic processes by introducing latent topic space. For an arbitrary word w in document d, (1) a topic z is first sampled from the multinomial distribution θ_d, which is generated from the Dirichlet prior parameterized by α; (2) and then the word w is generated from multinomial distribution ψ_z, which is generated from the Dirichlet prior parameterized by β. The two Dirichlet priors for documents-topic distribution θ_d and topic-word distribution ψ_z reduce the probability of overfitting training documents and enhance the ability of inferring topic distribution for new documents.

In cQA, the historical questions in the archives can be considered as documents. In this paper, we employ the state-of-the-art topic model −− LDA (Blei et al., 2003) to discover the latent topics in the content of questions. We assume that a queried question q and the historical questions Q in cQA archives are represented by a distribution over topics. We obtain the topic distribution of a question by merging the topic distributions of words in question. Formally, we have

$$P_{TM}(z|q) = \frac{1}{|q|}(\lambda_1 \sum_{w \in q} P(z|w) \tag{7}$$

Then, we assume that a question Q in the archives and a queried question q have the same prior probability, so the score function between the two questions can be written as:

$$\begin{aligned} P_{TM}(q|Q) &= \sum_z P(q|z) P_{TM}(z|Q) \\ &= \sum_{z \in K} \frac{P(z|q)P(q)}{p(z)} P_{TM}(z|Q) \\ &= \frac{K}{|q|} \sum_{z \in K} P_{TM}(z|q) P_{TM}(z|Q) \end{aligned} \tag{8}$$

4.2 Topic Model Incorporated with Category Information

In cQA, the questions are organized into a hierarchy of categories. For example, the subcategory "Computer Networking" is a child category of "Computers & Internet" in Yahoo! Answers. When a user asks a question, the user chooses a category for the question and at then post the question in that category. For example, the questions in the subcategory "Computer Networking" mainly related to computer software or networking equipments.

To utilize the category information provided by cQA, we propose a topic model incorporated with category information (TMC) to discover the latent topics in the content of questions. The graphic

representation of our proposed TMC model is presented in Figure 3. Inspired by the related work on topic analysis (Blei et al., 2003; Griffiths and Steyvers, 2004; Zhou et al., 2008; Wang and McCallum, 2006; Guo et al., 2008; Celikyilmaz et al., 2010; Jo and Oh, 2011), we make the following assumptions about the probabilistic structure of TMC model. First, each question is modeled as a multinomial distribution over latent topics, and each topic is modeled as a multinomial distribution over words and a multinomial distribution over categories. Second, the prior distributions for topics, words and categories follow different parameterized Dirichlet distribution, which is conjugate prior for multinomial distribution. In Figure 3, for each word w in question q, a topic z is first drawn from the multinomial distribution θ_q, and then a word is sampled from the multinomial distribution ϕ_z and a category c is also sampled from the multinomial distribution ψ_z for the word. Repeating this process N_q times, we get the words and category for a question. We obtain the whole question set by repeating the above process N times. After that, we obtain the topic distribution of a question by merging the topic distributions of words category. So equation (7) can be rewritten as:

$$P_{TMC}(z|q) = \frac{1}{1+|q|}(\lambda_2 P(z|c) + \lambda_3 \sum_{w \in q} P(z|w)) \quad (9)$$

In equation (9), the topic distribution of question category is modeled by $\lambda_2 P(z|c)$, the topic distribution of words in question is modeled by $\lambda_3 \sum_{w \in q} P(z|w)$. The relative importance of these two parts is adjusted through λ_2 and λ_3.

Introducing the category information into the process of discovering the latent topics, equation (8) can be rewritten as:

$$\begin{aligned} P_{TMC}(Q|q) &= \sum_z P(Q|z) P_{TMC}(z|q) \\ &= \sum_{z \in K} \frac{P(z|q)P(q)}{p(z)} P_{TMC}(z|Q) \\ &= \frac{K}{|q|} \sum_{z \in K} P_{TMC}(z|q) P_{TMC}(z|Q) \quad (10) \end{aligned}$$

4.3 Parameter Estimation for TMC

After introducing our proposed TMC method, we will describe how to estimate the parameter used in the model. In TMC, we introduce the new parameters, which lead to the inference not be done exactly. Expectation-Maximum (EM) algorithm is a possible choice for estimating the parameters of models with latent variables. However, EM suffers from the possibility of running into local maxima and the high computational burden. Therefore, we employ an alternative approach − Gibbs sampling (Griffiths, 2002), which is gaining popularity in recent work on latent topic analysis (Griffiths and Steyvers, 2004; Zhou et al., 2008; Wang and McCallum, 2006; Guo et al., 2008; Jo and Oh, 2011).

After training the model, we can get the following parameter estimations as:

$$\hat{\theta}_{qz} = \frac{n_{qz} + \alpha_z - 1}{\sum_{z'=1}^{K}(n_{qz'} + \alpha_{z'}) - 1}$$

$$\hat{\phi}_{zw} = \frac{n_{zw} + \beta_w - 1}{\sum_{v=1}^{|V|}(n_{zv} + \beta_v) - 1}$$

$$\hat{\psi}_{zc} = \frac{n_{zc} + \gamma_c - 1}{\sum_{c'=1}^{|C|}(n_{zc'} + \gamma_{c'}) - 1}$$

4.4 Combining the TMC with the TRLM for Question Retrieval

Since the TMC model and the translation-based language model use different strategies for question retrieval, it is interesting to explore how to combine their strength. In this section, we propose an approach to linearly combine the TMC model with the TRLM model for question retrieval. In this paper, we choose translation-based language model (TRLM) (Xue et al., 2008) as the foundation of our solution since TRLM has gained the state-of-the-art performance for question retrieval (Xue et al., 2008; Cao et al., 2010). Formally, we have

$$\begin{aligned} P_{TMC-TRLM}(q|Q) = \mu P_{TRLM}(q|Q) \\ + (1-\mu) P_{TMC}(q|Q) \quad (11) \end{aligned}$$

In equation (11), the relative importance of TMC and the TRLM is adjusted through μ. When $\mu = 1$, the retrieval model is based on TMC. When $\mu = 0$, the retrieval model is based on TRLM.

5 Experiments

5.1 Data Set and Evaluation Metrics

We collect the questions from Yahoo! Answers and use the *getByCategory* function provided in Yahoo! Answers API[3] to obtain Q&A threads

[3] http://developer.yahoo.com/answers

Category	#Size	Category	# Size
Arts & Humanities	86,744	Home & Garden	35,029
Business & Finance	105,453	Beauty & Style	37,350
Cars & Transportation	145,515	Pet	54,158
Education & Reference	80,782	Travel	305,283
Entertainment & Music	152,769	Health	132,716
Family & Relationships	34,743	Sports	214,317
Politics & Government	59,787	Social Science	46,415
Pregnancy & Parenting	43,103	Ding out	46,933
Science & Mathematics	89,856	Food & Drink	45,055
Computers & Internet	90,546	News & Events	20,300
Games & Recreation	53,458	Environment	21,276
Consumer Electronics	90,553	Local Businesses	51,551
Society & Culture	94,470	Yahoo! Products	150,445

Table 2: Number of questions in each first-level category

from the Yahoo! site. More specifically, we utilize the *resolved* questions and the resulting question repository that we use for question retrieval contains 2,288,607 questions. Each resolved question consists of four parts: "question title", "question description", "question answers" and "question category". For question retrieval, we only use the "question title" part and "question category" part. It is assumed that the titles and categories of the questions already provide enough semantic information. There are 26 categories at the first level and 1,262 categories at the leaf level. Each question belongs to a unique leaf category. Table 2 shows the distribution across first-level categories of the questions in the training data set. To learn the translation probabilities, we use about one million question-answer pairs from another data set.[4]

We randomly select 252 questions for test set and another 252 questions for development set. We select the test set and development set in proportion to the number of questions and categories against the whole distribution to have a better control over a possible imbalance. To obtain the ground-truth of question retrieval, we employ the Vector Space Model (VSM) (Salton et al., 1975) to retrieve the top 20 results and obtain manual judgements. The top 20 results don't include the queried question itself. Given a returned result by VSM, an annotator is asked to label it with "relevant" or "irrelevant". If a returned result is considered semantically equivalent to the queried question, the annotator will label it as "relevant"; otherwise, the annotator will label it as "irrelevant". Two annotators are involved in the annotation process. If a conflict happens, a third person will make judgement for the final result. In the process of manually judging questions, the annotators are presented only the questions. **Metrics**: We evaluate the performance of our approach using the following metrics: **Mean Average Precision** (MAP) and **Precision@n** (P@n). MAP rewards methods that return relevant questions early and also rewards correct ranking of the results. P@n reports the fraction of the top-n questions retrieved that are relevant. We perform a significant test, i.e., a t-test with a default significant level of 0.05.

Parameter Selection: The experiments use many parameters. Following the literature, we set the smoothing parameter λ in equations (1), (3) and (5) to 0.2 (Cao et al., 2010), and the parameter δ in equation (6) to 0.8 (Xue et al., 2008; Cao et al., 2010), which controls the translation component's impact. Other parameters are tuned on the development set, as we will show in the experiments.

5.2 Topic Number Selection

In this section, we concentrate on how to select proper topic numbers to obtain our model with best performance on our test set and enough iterations in Algorithm 1 to prevent overfitting problem. Here, following (Guo et al., 2008), we use perplexity to estimate the performance of our model. We calculate the perplexity on development set, which is a sequence of tuples $(q, w, c) \in D_{dev}$:

$$\text{Perplexity}(D_{dev}) = exp\{-\frac{\sum_{(q,w,c)\in D_{dev}} \ln P(w,c|q)}{|D_{dev}|}\}$$

Here, the probability $P(w, c|q)$ is calculated according to the parameters trained from the historical question-answer pairs:

$$P(w,c|q) = \sum_{z=1}^{K} P(w|z)P(c|z)P(z|q)$$

Figure 4(a) shows the influence of iteration number of Gibbs sampling on the model generalization ability. Empirically, we set the topic number as 100 and change the iteration number in the experiments. Note that the lower perplexity value indicates better generalization ability on the hold-out testing set. From Figure 4(a), it is seen that the perplexity values decreases dramatically when the iteration times are below 200.

Figure 4(b) shows the perplexity values for different settings of topic number. From the Figure, we see that the perplexity decreases when the

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

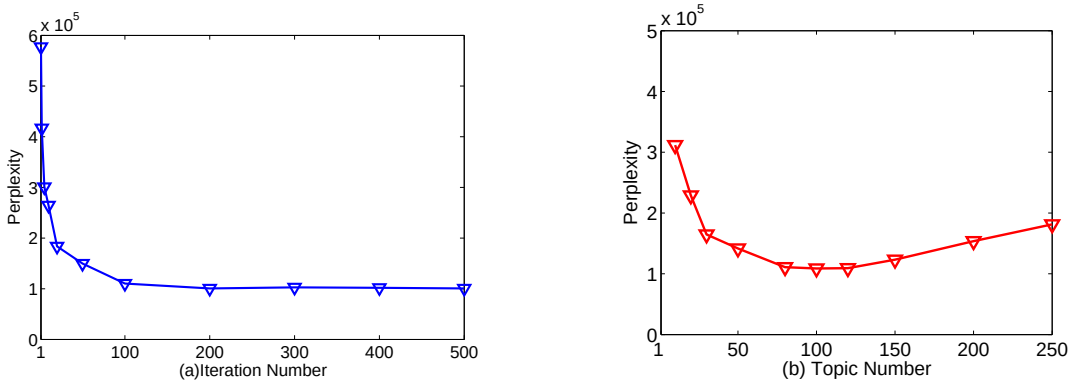

Figure 4: Perplexity on different iteration numbers(a) and topic number selection(b).

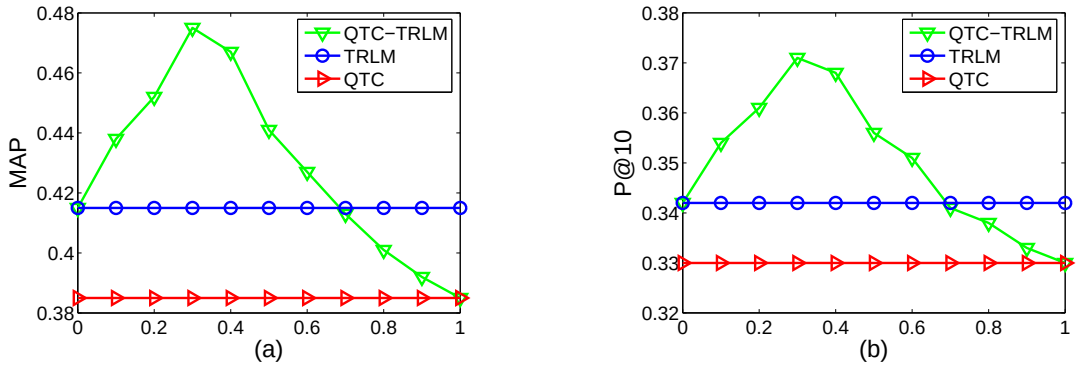

Figure 5: The relative importance of μ on the performance of TMC-TRLM.

number of topics starts to increase. However, after a certain point, the perplexity values start to increase. Based on the above experiments, we train our model using 100 topics and 200 iteration times.

5.3 The Relative Importance of Parameter μ

In equation (11), we use the parameter μ to adjust the relative importance of the TMC and the TRLM. Figure 5 illustrates the relative importance the value of μ is on the performance of question retrieval in terms of MAP and P@10, respectively. The TMC and TRLM are used for reference. The results are obtained with the 252 questions on the development set. From Figure 5, we see that for MAP and P@10, the combined model TMC-TRLM performs better than the TMC and TRLM when μ is between 0 and 0.7. In both cases, a relatively broad set for good parameter values is observed.

5.4 The Effectiveness of Our Proposed TMC Model

Table 3 shows the main results of question retrieval using the baseline methods and our proposed TMC-TRLM. In Table 3, VSM refers to the vector space model of (Salton et al., 1975); BM25 refers to the model of (Robertson et al., 1994); LM refers to the language model of (Zhai and Lafferty, 2001); TR refers to the translation model of (Jeon et al., 2005; Xue et al., 2008), TRLM refers to the translation-based language model of (Xue et al., 2008) and TRLM+CE refers to the method of (Cao et al., 2010).[5] In row 7, we show our approach and choose the best parameter $K = 100$. There are some clear trends in the results of Table 3:

#	Models	MAP	P@10
1	VSM	0.242	0.226
2	BM25	0.301	0.294
3	LM	0.352	0.327
4	TR	0.383	0.330
5	TRLM	0.415	0.342
6	TRLM+CE	0.437	0.358
7	TMC	0.385	0.331
8	**TMC-TRLM** ($K = 100$)	**0.475**	**0.371**

Table 3: Comparison with different methods for question retrieval.

[5]Here, we implement the method of (Cao et al., 2010) and use the TRLM to compute the global relevance and local relevance.

(1) The simple unigram language model (LM) performs slightly better than the classical retrieval models: VSM and BM25 (row 1 vs. row 3; row 2 vs. row 3).

(2) Translation model (TR) outperforms the LM by significant margins (row 3 vs. row 4).

(3) Translation-based language model (TRLM) significantly outperforms the translation model (TR) (row 4 vs. row 5), similar observations have been done by Xue et al. (2008).

(4) Exploiting category information of questions into the translation-based language model (TRLM) can significantly improve the question retrieval performance (row 5 vs. row 6), similar observations have been done by Cao et al. (2010).

(5) Our proposed approach TMC does not outperform the baseline methods TRLM and TRLM+CE (row 5 vs. row 7; row 6 vs. row 7). This demonstrates that the knowledge extracted from TMC is not as effective as that extracted from TRLM for question retrieval. TRLM learns the word-to-word translation probabilities from parallel corpus collected from question answer archives. However, TMC models word-category-topic distribution from the whole question answer content. The knowledge extracted from TMC is much noisier than that of TRLM. We suspect the above reason leads to the poor performance of TMC.

(6) Our proposed approach TMC-TRLM significantly outperforms the baseline methods TRLM and TRLM+CE (row 5 vs. row 8; row 6 vs. row 8). We conduct a significant test (t-test) on the improvements of our approach over TRLM and TRLM+CE. The result indicates that the improvements are statistically significant in terms of all the evaluation measures.[6] This demonstrates that the knowledge extracted from TMC is complementary to the knowledge extracted from TRLM+CE for question retrieval.

5.5 The Effectiveness of Category Information

Like the previous approaches, we treat the questions as a multinomial distribution over latent topics, and each topic is a multinomial distribution over words too. Different from previous work on topic analysis (Blei et al., 2003; Griffiths and Steyvers, 2004; Zhou et al., 2008; Wang and McCallum, 2006; Guo et al., 2008; Celikyilmaz et al., 2010; Jo and Oh, 2011), we introduce the category information of questions, which is predefined by cQA services, into the process of discovering latent topics. To see how much the category information benefit the question retrieval, we introduce a baseline method for comparison. The baseline method (denoted as TM-TRLM) is used to denote the proposed method without using the category information. Table 5 provides the comparison. From the Table, we see that the exploring category information can significantly improve the performance for question retrieval (row 1 vs. row 2).

[6]The comparisons are significant at $p < 0.05$.

#	Models	MAP	P@10
1	TM-TRLM	0.454	0.366
2	**TMC-TRLM**	**0.475**	**0.371**

Table 4: The effectiveness of category information for question retrieval.

6 Conclusions and Future Work

In this paper, we present a new approach to discover the latent topic of questions for improving the performance of translation-based language model for question retrieval. Experiments conducted on real cQA data demonstrate that our proposed approach significantly outperforms the state-of-the-art methods (TRLM and TRLM+CE).

There are some ways in which this research could be continued. First, question structure should be considered, so it is necessary to combine the proposed approach with other question retrieval methods (e.g., (Duan et al., 2008; Wang et al., 2009; Bunescu and Huang, 2010)) to further improve the performance. Second, we will try to investigate the use of the proposed approach for other kinds of data set, such as categorized questions from forum sites and FAQ sites.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 60875041 and No. 61070106). We thank the anonymous reviewers for their insightful comments.

References

A. Berger, R. Caruana, D. Cohn, D. Freitag, and V. Mittal. 2000. Bridging the lexical chasm: statistical approach to answer-finding. In *Proceedings of SIGIR*, pages 192-199.

A. Berger and J. Lafferty. 1999. Information retrieval as statistical translation. In *Proceedings of SIGIR*, pages 222-229.

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

D. M. Blei, A. Ng, and M. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993-1022.

P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263-311.

R. Bunescu and Y. Huang. 2010. Learing the relative usefulness of questions in community QA. In *Proceedings of EMNLP*, pages 97-107.

X. Cao, G. Cong, B. Cui, C. S. Jensen, and C. Zhang. 2009. The use of categorization information in language models for question retrieval. In *Proceedings of CIKM*.

X. Cao, G. Cong, B. Cui, and C. S. Jensen. 2010. A generalized framework of exploring category information for question retrieval in community question answer archives. In *Proceedings of WWW*.

A. Celikyilmaz, D. Hakkani-Tur, and G. Tur. 2010. LDA based similarity modeling for question answering. In *Proceedings of ACL*.

H. Duan, Y. Cao, C. Y. Lin, and Y. Yu. 2008. Searching questions by identifying questions topics and question focus. In *Proceedings of ACL*, pages 156-164.

T. Griffiths. Gibbs sampling in the generative model of latent dirichlet allocation. http://www-psych.stanford.edu/ gruffydd/cogsci02/lda.ps.

T. Griffiths and M. Steyvers. 2004. Finding scientific topics. In National Academy of Sciences.

J. Guo, S. Xu, S. Bao, and Y. Yu. 2008. Tapping on the potential of Q&A community by recommending answer providers. In *Proceedings of CIKM*.

J. Jeon, W. Bruce Croft, and J. H. Lee. 2005. Finding similar questions in large question and answer archives. In *Proceedings of CIKM*, pages 84-90.

Y. Jo and A. Oh. 2011. Aspect and sentiment unification model for online review analysis. In *Proceedings of WSDM*.

J. -T. Lee, S. -B. Kim, Y. -I. Song, and H. -C. Rim. 2008. Bridge lexical gaps between queries and questions on large online Q&A collections with compact translation models. In *Proceedings of EMNLP*, pages 410-418.

J. M. Ponte and W. B. Croft. 1998. A language modeling approach to information retrieval. In *Proceedings of SIGIR*.

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

G. Salton, A. Wong, and C. S. Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613-620.

X. Wang and A. McCallum. 2006. Topic over time: a non-markov conditionals-time model of topical trends. In *Proceedings of SIGKDD*, pages 424-433.

K. Wang, Z. Ming, and T-S. Chua. 2009. A syntactic tree matching approach to finding similar questions in community-based qa services. In *Proceedings of SIGIR*, pages 187-194.

X. Xue, J. Jeon, and W. B. Croft. 2008. Retrieval models for question and answer archives. In *Proceedings of SIGIR*, pages 475-482.

C. Zhai and J. Lafferty. 2001. A study of smooth methods for language models applied to ad hoc information retrieval. In *Proceedings of SIGIR*, pages 334-342.

D. Zhou, J. Bian, S. Zheng, H. Zha, and C. L. Giles. 2008. Exploring social annotation for information retrieval. In *Proceedings of WWW*, pages 715-724.

G. Zhou, L. Cai, J. Zhao, and K. Liu. 2011. Phrase-based translation model for question retrieval in community question answer archives. In *Proceedings of ACL-HLT*, pages 653-662.