Question-Answer Topic Model for Question Retrieval in Community Question Answering

Zongcheng Ji[1,2], Fei Xu[1,2], Bin Wang[1] and Ben He[2]
[1]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[2]Graduate University of Chinese Academy of Sciences, Beijing, China
{jizongcheng, feixu1966}@gmail.com, wangbin@ict.ac.cn, benhe@gucas.ac.cn

ABSTRACT
The major challenge for Question Retrieval (QR) in Community Question Answering (CQA) is the lexical gap between the queried question and the historical questions. This paper proposes a novel Question-Answer Topic Model (QATM) 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. Experiments conducted on a real world CQA dataset from Yahoo! Answers show that combining both parts properly can get more knowledge than each part or both parts in a simple mixing way and combining our QATM with the state-of-the-art translation-based language model, where the topic and translation information is learned from the question-answer pairs at two different grained semantic levels respectively, can significantly improve the QR performance.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords
Community Question Answering, Question-Answer Topic Model, Topic Model, Translation Model, Question Retrieval

1. INTRODUCTION
Community Question Answering (CQA) services have accumulated large archives of question-answer pairs. To reuse the invaluable resources, it is essential to develop effective retrieval models to retrieve similar questions, which are semantically equivalent or relevant to the queried questions.

The major challenge for question retrieval, as for most information retrieval tasks, is the *lexical gap* between the queried question and the historical questions in the CQA archives.

Most of previous work to bridge the lexical gap is based on the statistical translation approach [7, 9, 13], which learns the word-to-word translation probabilities from the historical comparable question-answer pairs *at the fine-grained lexical semantic level* (aka at the word level), while ignoring the latent semantic information in calculating the semantic similarity between the queried question and the historical questions.

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

Recently, some work has been proposed to use the latent semantic information to bridge the lexical gap in question retrieval [1, 15]. However, both of these work learns the semantic representation only from the question part (seeing a question as a common document), ignoring the important paired answer part. Furthermore, simply applying the existing methods to both parts can benefit from the answer part but only slightly as we have shown.

In this paper, we argue that it is beneficial to involve the answer part into learning the latent semantic topics for question retrieval. Thus, our research goal is to investigate how to develop effective topic models to learn the latent semantic topics from the question-answer pairs more precisely and effectively, and apply the models to improve the retrieval performance.

Specifically, we first propose a novel Question-Answer Topic Model (QATM) to model the question-answer relationships to learn the latent topics aligned across the historical comparable question-answer pairs *at the coarse-grained latent semantic level* to alleviate the lexical gap problem, with the assumption that a question and its paired answer share the same topic distribution. Second, realizing that the shared topic vector can be easily dominated by the longer one of the question-answer pair, we further extend QATM with posterior regularization [4] (QATM-PR) by constraining the question-answer pair have similar fractions of words assigned to each topic. Third, we introduce the retrieval model using QATM, which ranks historical questions by their probabilities of generating a queried question. Then, we propose a general framework, which combines the translation and topic information learned from the question-answer pairs at two different grained semantic levels, for question retrieval in CQA. Finally, we conduct experiments on a real world CQA dataset from Yahoo! Answers. We proceed to present them in the following sections.

2. PROPOSED APPROACH
In a CQA archive, since the asker and answerer may express similar meanings with different words, it is natural to use the question-answer pairs as the parallel corpus that is used for estimating the translation probabilities, which can be seen as the information learned at the fine-grained lexical semantic level. Similarly, inspired by the work of [5, 11], the question from the asker and the answer from the answerer can also be assumed to be written in two different languages, while sharing as much as possible the same topic fractions. Based on this assumption, we propose a model called Question-Answer Topic Model (QATM) to model the question-answer relationships at the coarse-grained latent semantic level to learn the latent semantic topics aligned across the question-answer pairs.

2.1 Question-Answer Topic Model

We assume that a question $\mathbf{q} = w_{\mathbf{q}_1} \dots w_{\mathbf{q}_{|\mathbf{q}|}}$ and its paired answer $\mathbf{a} = w_{\mathbf{a}_1} \dots w_{\mathbf{a}_{|\mathbf{a}|}}$ share the same topic distribution, but use different (perhaps overlapping) vocabularies to express these topics. Figure 1 shows the graphical representation of QATM, while Figure 2 summarizes the generative story of generating a question-answer pair.

Thus, the log-likelihood of a whole collection of question-answer pairs $C = \{(\mathbf{q}, \mathbf{a}) | \mathbf{a} \text{ is the answer of } \mathbf{q}\}$ together with the parameters $\Lambda = \{\theta, \Phi_z^{\mathbf{q}}, \Phi_z^{\mathbf{a}}\}$ is

$$\log P(C|\Lambda) = \log\left(P(\Phi_z^{\mathbf{q}}|\beta^{\mathbf{q}})P(\Phi_z^{\mathbf{a}}|\beta^{\mathbf{a}})\prod_{(\mathbf{q},\mathbf{a})}P(\theta|\alpha)P((\mathbf{q},\mathbf{a})|\Lambda)\right) \quad (1)$$

where

$$P((\mathbf{q},\mathbf{a})|\Lambda) = \prod_{w_{\mathbf{q}}\in\mathbf{q}}\sum_z P(w_{\mathbf{q}}|\Phi_z^{\mathbf{q}})P(z|\theta) \cdot \prod_{w_{\mathbf{a}}\in\mathbf{a}}\sum_z P(w_{\mathbf{a}}|\Phi_z^{\mathbf{a}})P(z|\theta) \quad (2)$$

In our work, we learn the parameters using MAP estimation, treating α, $\beta^{\mathbf{q}}$, $\beta^{\mathbf{a}}$ as hyper-parameters, each corresponding to one Dirichlet prior. According to the standard EM algorithm, we can use the iterative updating formulas given in Figure 3 to estimate all the parameters.

2.2 Posterior Regularization

A question-answer pair is expected to not only share the same prior distribution over topics, but also contain similar fractions of words assigned to each topic. Since MAP estimation of the shared topic vector $P(z|\theta)$ (in the M-Step of Figure 3) is concerned with explaining the union of words in the question and its paired answer, and can be easily dominated by the longer one of the two, it does not guarantee that each topic z occurs with similar frequency in the question and its paired answer. Thus, following [5, 11], we also extend QATM with posterior regularization [4] (QATM-PR) by constraining the question-answer pair have similar fractions of words assigned to each topic, trained using a modified E-Step in the original EM algorithm.

2.3 Ranking Historical Questions

Since the Question-Answer Topic Model and the Translation-based Methods cover different grained semantic levels, it is interesting to explore how to combine their strength. In this paper, we choose the Translation-based Language Model (TransLM) as the Translation-based Method since TransLM has gained the state-of-the-art performance for question retrieval [3, 13]. Finally, we have the following ranking function for question retrieval:

$$P_{QATM+TransLM}(\mathbf{q}|Q) = (1-\mu)P_{QATM}(\mathbf{q}|Q) + \mu P_{TransLM}(\mathbf{q}|Q) \quad (3)$$

$$P_{QATM}(\mathbf{q}|Q) = \prod_{w\in\mathbf{q}} P_{qatm}(w|Q) \quad (4)$$

$$P_{qatm}(w|Q) = \sum_z P(w|\Phi_z^Q)P(z|\theta^{(Q,A)}) = \sum_z P(w|\Phi_z^Q)P(z|\theta^{(Q)}) \quad (5)$$

$$P_{TransLM}(\mathbf{q}|Q) = \prod_{w\in\mathbf{q}}(1-\lambda)P_{mx}(w|Q) + \lambda P_{ml}(w|C) \quad (6)$$

$$P_{mx}(w|Q) = (1-\gamma)P_{ml}(w|Q) + \gamma P_{trans}(w|Q) \quad (7)$$

$$P_{trans}(w|Q) = \sum_{t\in Q} T(w|t)P_{ml}(t|Q) \quad (8)$$

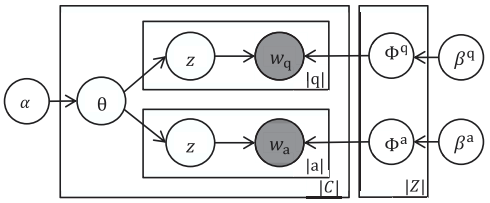

Figure 1: Graphical representation of QATM

For each topic z
  Choose a pair of different topic-specific word distributions $(\Phi_z^{\mathbf{q}}, \Phi_z^{\mathbf{a}})$:
    A topic-specific question-word distribution $\Phi_z^{\mathbf{q}} \sim Dirichlet(\beta^{\mathbf{q}})$.
    A topic-specific answer-word distribution $\Phi_z^{\mathbf{a}} \sim Dirichlet(\beta^{\mathbf{a}})$.

For each question-answer pair $(\mathbf{q}, \mathbf{a})$
  Choose a topic distribution $\theta \sim Dirichlet(\alpha)$.
  For each word position i in $\mathbf{q}$
    Choose a topic $z \sim Multinomial(\theta)$.
    Choose a word $w_{\mathbf{q}} \sim Multinomial(\Phi_z^{\mathbf{q}})$.
  For each word position j in $\mathbf{a}$
    Choose a topic $z \sim Multinomial(\theta)$.
    Choose a word $w_{\mathbf{a}} \sim Multinomial(\Phi_z^{\mathbf{a}})$.

Figure 2: Generative story for QATM

E-Step:

$$P(z|\mathbf{q}, w_{\mathbf{q}}) = \frac{P(w_{\mathbf{q}}|\Phi_z^{\mathbf{q}})P(z|\theta)}{\sum_{z'}P(w_{\mathbf{q}}|\Phi_{z'}^{\mathbf{q}})P(z'|\theta)}$$

$$P(z|\mathbf{a}, w_{\mathbf{a}}) = \frac{P(w_{\mathbf{a}}|\Phi_z^{\mathbf{a}})P(z|\theta)}{\sum_{z'}P(w_{\mathbf{a}}|\Phi_{z'}^{\mathbf{a}})P(z'|\theta)}$$

M-Step:

$$P(w_{\mathbf{q}}|\Phi_z^{\mathbf{q}}) = \frac{\sum_{w_{\mathbf{q}}} n(\mathbf{q}, w_{\mathbf{q}})P(z|\mathbf{q}, w_{\mathbf{q}}) + (\beta^{\mathbf{q}} - 1)}{\sum_{\mathbf{q}}\sum_{w_{\mathbf{q}}'} n(\mathbf{q}, w_{\mathbf{q}}')P(z|\mathbf{q}, w_{\mathbf{q}}') + |\mathbf{q}|(\beta^{\mathbf{q}} - 1)}$$

$$P(w_{\mathbf{a}}|\Phi_z^{\mathbf{a}}) = \frac{\sum_{w_{\mathbf{a}}} n(\mathbf{a}, w_{\mathbf{a}})P(z|\mathbf{a}, w_{\mathbf{a}}) + (\beta^{\mathbf{a}} - 1)}{\sum_{\mathbf{a}}\sum_{w_{\mathbf{a}}'} n(\mathbf{a}, w_{\mathbf{a}}')P(z|\mathbf{a}, w_{\mathbf{a}}') + |\mathbf{a}|(\beta^{\mathbf{a}} - 1)}$$

$$P(z|\theta) = \frac{\sum_{w_{\mathbf{q}}} n(\mathbf{q}, w_{\mathbf{q}})P(z|\mathbf{q}, w_{\mathbf{q}}) + \sum_{w_{\mathbf{a}}} n(\mathbf{a}, w_{\mathbf{a}})P(z|\mathbf{a}, w_{\mathbf{a}}) + (\alpha - 1)}{\sum_{z'}\left(\sum_{w_{\mathbf{q}}} n(\mathbf{q}, w_{\mathbf{q}})P(z'|\mathbf{q}, w_{\mathbf{q}}) + \sum_{w_{\mathbf{a}}} n(\mathbf{a}, w_{\mathbf{a}})P(z'|\mathbf{a}, w_{\mathbf{a}})\right) + |Z|(\alpha - 1)}$$

Figure 3: Iterative updating formulas for estimating parameters $\Lambda = \{\theta, \Phi_z^{\mathbf{q}}, \Phi_z^{\mathbf{a}}\}$ of QATM. Here, $n(\mathbf{q}, w_{\mathbf{q}})$ is the frequency of term $w_{\mathbf{q}}$ in question $\mathbf{q}$, and $n(\mathbf{a}, w_{\mathbf{a}})$ is the frequency of term $w_{\mathbf{a}}$ in answer $\mathbf{a}$. $|\mathbf{q}|$, $|\mathbf{a}|$ and $|Z|$ are the size of question $\mathbf{q}$, answer $\mathbf{a}$ and topics Z, respectively.

Here, (Q, A) denotes the historical question-answer pair. $P(z|\theta^{(Q,A)})$ and $P(w|\Phi_z^Q)$ are the topic distribution for each historical question-answer pair and the topic-specific question-word distribution, respectively; and μ is the parameter to balance between QATM and TransLM. $P_{ml}(w|Q)$ and $P_{ml}(w|C)$ are the maximum likelihood estimation of word w in Q and the whole collection C, respectively; and λ is the Jelinek-Mercer smoothing factor. $T(w|t)$ is the probability of translating a word t in Q into a word w in $\mathbf{q}$. γ is the parameter to balance between the Language Model and the Translation Model.

Notice that, although the query term w in Equation (5) is generated from the topic-specific word distributions in *question language*, the topic distribution in QATM is learned from the parallel question-answer pairs. In addition, $\sum_z P(w|\Phi_z^Q)P(z|\theta^{(Q)})$ in Equation (5) can be seen as performing a generative process of a queried question term w from the *question language*. Thus, the question retrieval funciton using our QATM in Equation (4) can be seen as ranking a historical question Q by its probabilities of generating a queried question $\mathbf{q}$.

Table 1: Comparison of different models for question retrieval. Row 1 ~ Row 8 are the results of different single models. Row 9 ~ Row 13 are the results of incorporating five different topic models with LM. Row 14 ~ Row 18 are the results of incorporating five different topic models with TransLM. "*" and "+" indicate statistically significant improvements ($p < 0.05$ using a paired *t*-test) over LM and TransLM, respectively.

#	Models	MAP	P@10	#	Models	MAP	P@10
1	LM	0.3067	0.2430	9	PLSA_q+LM	0.3070	0.2365
2	Trans	0.3327*	0.2565*	10	PLSA_a+LM	0.2911	0.2330
3	TransLM	0.3634*	0.2700*	11	PLSA_qa+LM	0.3169*	0.2430*
-	---	---	---	12	QATM+LM	0.3631*	0.2630*
-	---	---	---	13	QATM-PR+LM	0.3682*+	0.2630*
4	PLSA_q	0.2581	0.1930	14	PLSA_q+TransLM	0.3675*+	0.2730*+
5	PLSA_a	0.2531	0.1830	15	PLSA_a+TransLM	0.3549*	0.2665*
6	PLSA_qa	0.2599	0.1965	16	PLSA_qa+TransLM	0.3684*+	0.2800*+
7	QATM	0.3126*	0.2530*	17	QATM+TransLM	0.3977*+	0.2930*+
8	QATM-PR	0.3186*	0.2530*	18	QATM-PR+TransLM	**0.4026***+	**0.3000***+

3. EXPERIMENTS

3.1 Dataset

We collect all the questions-answer pairs from Yahoo! Answers using *getByCategory* function provided in Yahoo! Answers API[1]. Specifically, we utilize the *resolved* questions under the top-level category, namely "Computers & Internet". The resulting *question repository* that we use for question retrieval contains 524,195 questions. We use 1,186,853 question-answer pairs from another dataset[2] for training the word translation probabilities and the topic models. With the above two dataset, we use the *subject* field as the question part and the *bestanswer* field as the answer part. For preprocessing, all the questions and answers are lowercased and stopwords are removed using a standard list of 418 words.

To create the *test set* for question retrieval, we randomly select 200 questions from the question repository, with the remaining 523,995 question-answer pairs as the retrieval corpora. To create the *ground-truth* for question retrieval, we obtain annotation by *pooling* the top 20 results from all the models for each query.

3.2 Question Retrieval Models

Baselines: Three types of baselines are used to compare with our proposed retrieval models, which are summarized as follows:

(1) Language Model (LM)

(2) Translation Model (Trans) and the state-of-the-art Translation-Based Language Model (TransLM). We use GIZA++[3] to train the IBM model 1 to get the word translation probabilities.

(3) Topic Models learned from the question part (PLSA_q), the answer part (PLSA_a) and the question-answer pairs in a simple "plus" way[4] (PLSA_qa) with the traditional topic model PLSA [6]. We use folding-in with 10 EM iterations to map each question in the retrieval corpora to its corresponding topic vector. We use 200 topics for all the topic models including our proposed QATM.

We also combine all the topic models with LM and TransLM to see the interpolation performance. The name of the combined models is denoted using the symbol "+" to plus each name of the models. For example, QATM+TransLM means we will mix QATM and TransLM linearly.

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

[2]The Yahoo!Webscope dataset ydata-yanswers-all-questions-v1_0, available at http://research.yahoo.com/Academic_Relations

[3]http://www.fjoch.com/GIZA++.html

[4] Here, we simply "plus" the question and its paired answer as a whole text. It is different from our proposed QATM, which can be seen as in a "parallel" way.

Parameter Setting: The experiments use many parameters. Following the literature, we set the smoothing parameter $\lambda = 0.2$ [3] in Equations (6), $\gamma = 0.8$ [3, 13] in Equation (7) and hyper-parameters $\alpha = 1.1$, $\beta^{\mathbf{q}} = 1.01$, $\beta^{\mathbf{a}} = 1.01$ [5, 11] in Equation (1). Parameter μ in Equation (3) is tuned via 5-fold cross-validation: in each trial, we tune the parameters with four of the five subsets and then apply it to the remaining one. All the results reported above are those averaged over the five trials.

3.3 Experimental Results and Discussion

We evaluate the performance of all the ranking models using MAP and P@10. We also perform a significance test using a paired ***t***-test with a default significant level of 0.05.

3.3.1 The Effectiveness of the Proposed QATM

Table 1 presents the comparison of different models for question retrieval. There are some clear trends in the results of Table 1:

(1) Trans significantly outperforms LM (row 1 vs. row 2). TransLM significantly outperforms Trans (row 2 vs. row 3). The results are consistent with the former work [13].

(2) QATM outperforms LM significantly (row 1 vs. row 7).

(3) QATM does not outperform Trans (row 2 vs. row 7). We think that there are mainly two possible reasons: First, Trans is trained at the fine-grained lexical semantic level, which may contribute much to find similar questions than QATM which is trained at the coarse-grained latent semantic level. Second, from Equation (8), we can see that Trans includes LM. Specifically, when we assume $T(w|t) = 1$, Equation (8) will reduce to LM, but QATM is not the case. What excites us is that when interpolating them with LM, QATM+LM performs as well as TransLM with no significant difference (row 3 vs. row 12). This demonstrates that the topic information extracted from parallel question-answer pairs at the latent semantic level is in some sense as much as the translation information learned at the lexical semantic level.

(4) When further incorporating QATM with TransLM, the retrieval performance can be further improved (row 3 vs. row 17). This means that the topic information learned at the coarse-grained semantic level using our QATM can bring more knowledge to enhance the state-of-the-art TransLM, where the translation information is learned at the fine-grained lexical semantic level.

(5) When using posterior regularization (PR) to constrain the paired question and answer not only to share the same prior topic distribution, but also to have similar fractions of words assigned to each topic, the difference between QATM and QATM-PR is statistically significant (row 7 vs. row 8; row 12 vs. row 13; row 17 vs. row 18).

3.3.2 The Effectiveness of Different Topic Models

In this section, we will show how effective to learn the latent semantic topics jointly from both parts properly than each part or a simple "plus" way. There are mainly four ways (five models) to learn the latent topics. Three of the models are the third type of baselines. The other two are the proposed QATM and QATM-PR.

First, we do experiments only using the five topic models as the document model for question retrieval. The results in Table 1 suggest several observations. (1) Only using PLSA_q, PLSA_a and PLSA_qa as the document model hurts the ranking performance and is worse than LM (row 4~6 vs. row 1). (2) The latent topics learned in a simple "plus" way are better than that only from the question part or the answer part (row 4~5 vs. row 6). (3) QATM significantly outperforms LM (row 7 vs. row 1), and of course outperforms PLSA_q, PLSA_a and PLSA_qa (row 4~6 vs. row 7). This shows that QATM can learn the latent topics from the question-answer pairs more effectively and precisely. (4) The performance of QATM can be further improved by introducing constraints in the EM training to force the paired question and answer to share the same proportion of topics (row 7 vs. row 8).

Then, we do the comparison experiments of the five topic models incorporated with LM. The results in Table 1 show the following observations. (1) After linearly combined with LM, PLSA_q and PLSA_a do not improve LM (row 9~10 vs. row 1), however, PLSA_qa can improve LM significantly (row 11 vs. row 1). This suggests that learning the latent topics from the question-answer pairs is much better than only from the question part or the answer part. Similar observations can be seen in the first group of comparisons. (2) Incorporating QATM with LM can benefit LM significantly (row 12 vs. row 1). (3) QATM-PR benefits LM more than QATM (row 12 vs. row 13).

Finally, we also compare the five topic models incorporated with TransLM to see whether the topic models can bring more knowledge into the state-of-the-art TransLM. The observations from Table 1 are as follows. (1) PLSA_a does not improve TransLM (row 15 vs. row 3); however, PLSA_q and PLSA_qa can improve TransLM (row 14, row 16 vs. row 3). (2) Incorporating QATM or QATM-PR with TransLM outperforms TransLM significantly (row 17~18 vs. row 3). These results denote again that our proposed QATM can learn the latent topics from the question-answer pairs more effectively and precisely.

From the above three groups of comparisons, we may get the following conclusions.

(1) To learn the latent topics only from the answer part will not improve the retrieval performance.

(2) When incorporating the answer part with the question part, no matter in a simple "plus" way or in a new parallel way, we can learn more knowledge on the latent topics than either from the question part or from the answer part. This indicates that learning the latent topics from both parts together can benefit from the answer part, which is not investigated in previous work [1, 15].

(3) Our proposed QATM in a parallel way can benefit more from the answer part to learn the latent topics than the simple "plus" way by significant margins. The main reason maybe we only keep the topic-word distributions from the question part for question retrieval in Equation (5), however, PLSA_qa keeps the topic-word distributions in the whole text of question part and answer part, which will involve much noise. Thus, our proposed QATM can learn the latent topics more effectively and precisely.

4. RELATED WORK

Recently, question retrieval has been widely investigated in CQA data. Most of previous work focuses on the translation-based methods [7, 9, 13, 14] to alleviate the lexical gap problem. Besides, category information has been exploited [3, 2, 8, 10] in question retrieval task. Wang et al. [12] propose a syntactic tree matching method to find similar questions. Cai et al. [1], Zhou et al. [15] propose to learn the latent topics from the questions. Moreover, Cai et al. [1] also incorporate the question categories in learning the latent topics for further improvements.

5. CONCLUSION

In this paper, we propose a novel Question-Answer Topic Model (QATM) to learn the latent semantic topics from the question-answer pairs for question retrieval. Experiments conducted on a real world CQA dataset demonstrate that our proposed QATM significantly outperforms the other topic models learned from the question part, the answer part or both parts in a simple "plus" way with a traditional method. In addition, combining QATM with the state-of-the-art translation-based language model, where the topic and translation information is learned from the question-answer pairs at two different grained semantic levels respectively, can further improve the retrieval performance significantly.

There are some interesting future works to be continued. First, contexture information should be considered to learn the latent topics, since the phrase-based translation model [14] has been proposed to show the effectiveness of the contexture information. Second, category information can also be incorporated into our model for discovering latent topics in the context of questions.

6. ACKNOWLEDGMENTS

This work is supported by the National Science Foundation of China under Grant No. 61070111 and the Strategic Priority Research Program of the Chinese Academy of Sciences under Grant No. XDA06030200.

7. REFERENCES

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

[2] Xin Cao, Gao Cong, Bin Cui, Christian Søndergaard Jensen, and Ce Zhang. The use of categorization information in language models for question retrieval. In *CIKM*, pages 265-274, 2009.

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

[4] Kuzman Ganchev, João Graça, Jennifer Gillenwater, and Ben Taskar. Posterior regularization for structured latent variable models. *J. Mach. Learn. Res.*, **11**: p. 2001-2049, 2010.

[5] Jianfeng Gao, Kristina Toutanova, and Wen-tau Yih. Clickthrough-based latent semantic models for web search. In *SIGIR*, pages 675-684, 2011.

[6] Thomas Hofmann. Probabilistic latent semantic indexing. In *SIGIR*, pages 50-57, 1999.

[7] Jiwoon Jeon, W. Bruce Croft, and Joon Ho Lee. Finding similar questions in large question and answer archives. In *CIKM*, pages 84-90, 2005.

[8] Zongcheng Ji, Fei Xu, and Bin Wang. A category-integrated language model for question retrieval in community question answering. In *AIRS*, 2012.

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

[10] Zhao-Yan Ming, Tat-Seng Chua, and Gao Cong. Exploring domain-specific term weight in archived question search. In *CIKM*, pages 1605-1608, 2010.

[11] John C. Platt, Kristina Toutanova, and Wen-tau Yih. Translingual document representations from discriminative projections. In *EMNLP*, pages 251-261, 2010.

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

[13] Xiaobing Xue, Jiwoon Jeon, and W. Bruce Croft. Retrieval models for question and answer archives. In *SIGIR*, pages 475-482, 2008.

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

[15] Tom Chao Zhou, Chin-Yew Lin, Irwin King, Michael R. Lyu, Young-In Song, and Yunbo Cao. Learning to suggest questions in online forums. In *AAAI*, pages 1298-1303, 2011.