Question Retrieval with High Quality Answers in Community Question Answering

Kai Zhang*
State Key Laboratory of Software Development Environment
Beihang University, China
zhangkai@cse.buaa.edu.cn

Wei Wu
Microsoft Research
Beijing, China
wuwei@microsoft.com

Haocheng Wu
University of Science and Technology of China
Hefei, China
ustcwhc@outlook.com

Zhoujun Li
State Key Laboratory of Software Development Environment
Beihang University, China
lizj@buaa.edu.cn

Ming Zhou
Microsoft Research
Beijing, China
mingzhou@microsoft.com

ABSTRACT

This paper studies the problem of question retrieval in community question answering (CQA). To bridge lexical gaps in questions, which is regarded as the biggest challenge in retrieval, state-of-the-art methods learn translation models using answers under an assumption that they are parallel texts. In practice, however, questions and answers are far from "parallel". Indeed, they are heterogeneous for both the literal level and user behaviors. There are a particularly large number of low quality answers, to which the performance of translation models is vulnerable. To address these problems, we propose a supervised question-answer topic modeling approach. The approach assumes that questions and answers share some common latent topics and are generated in a "question language" and "answer language" respectively following the topics. The topics also determine an answer quality signal. Compared with translation models, our approach not only comprehensively models user behaviors on CQA portals, but also highlights the instinctive heterogeneity of questions and answers. More importantly, it takes answer quality into account and performs robustly against noise in answers. With the topic modeling approach, we propose a topic-based language model, which matches questions not only on a term level but also on a topic level. We conducted experiments on large scale data from Yahoo! Answers and Baidu Knows. Experimental results show that the proposed model can significantly outperform state-of-the-art retrieval models in CQA.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval models*

Keywords

comunity question answering; question retrieval; supervised topic model; answer quality

1. INTRODUCTION

Community question answering (CQA) portals, like Yahoo! Answers, Baidu Knows and Quora, have emerged as hot platforms for people to share their knowledge and learn from each other. In the last decade, these web sites have attracted a great number of users, and have accumulated a large amount of content generated by these users. The content is usually organized as questions and lists of answers associated with metadata like user votes to answers and askers' awards to the best answerers. This data has made CQA archives valuable repositories for various tasks like knowledge mining, question-answering and searching, etc.

One fundamental task for reusing content in CQA is finding similar questions for queries, as questions are the keys to accessing the knowledge in CQA. Many studies have been done along this line [6, 7, 9, 20]. One big challenge for question retrieval in CQA is that users are used to expressing similar meanings with different words, which creates lexical gaps when matching questions based on common terms. For example, we find that for a user query "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 have very similar meanings, since they share few common words, it is hard for classic retrieval models like BM25 [15] and language models for information retrieval ($LMIR$) [21] to recognize their similarity. The lexical gap has become a major barricade preventing traditional IR models from retrieving similar questions in CQA.

*The work is done when the author is an intern at Microsoft Research Asia.

CIKM '14, November 03 - 07 2014, Shanghai, China

http://dx.doi.org/10.1145/2661829.2661908.

Table 1: An example from Yahoo! Answers

Question (Question title)	Connected to the Internet but cannot get online!??
Description (Question body)	I have put a clean install of XP on my laptop and installed all the network and internet drivers from Dell.....
Answer 1	Does the Wi-Fi Connection have a passkey, because if it does the internet wont work if you have XP SP1. However you can upgrade your XP to SP2. Then it should work.Or you could take the passkey off your Wi-Fi
Answer 2	how have you posted this if you cant go on the internet

To overcome this issue, existing work considers leveraging answers of questions and learning translation models to improve traditional IR models. The basic assumption behind these models is that questions and answers are "parallel texts" and relationships of words can be established through word-to-word translation probabilities [20]. The translation models represent state-of-the-art methods for question retrieval in CQA. However, we argue that in practice, questions and answers are far from "parallel". Questions and answers are highly asymmetric on the information they contain. One question may have multiple answers. These answers may be diverse and contain much more information than the question. Moreover, the language of answers is usually more casual than that of questions. For example, answerers like using words such as 'hello', 'L.O.L', or 'wow' in their answers to make them friendly or appealing. These words do not help much in retrieving similar questions, since askers seldom use them. Additionally, on popular CQA sites like Yahoo! Answers, there is usually a large variance in quality of answers, from highly valuable knowledge to spam, and low quality answers take a relatively high proportion of all answers (more than 30% according to [10, 16]). Table 1 gives an example[1]. We can see that Answer 2 is a low quality answer regarding to the question. It makes no sense as an answer and is useless for bridging lexical gaps when retrieving the question. On the other hand, Answer 1, another answer in the answer list of the question, not only provides rich information for the question, but also has the potential to help question retrieval, as the answerer used 'wi-fi' as a related word of 'internet' in the context. Intuitively, low quality answers will pollute training data and make the learnt translation probabilities unreliable. It is better to select good answers for model learning. For translation-based methods, it is not trivial to do so.

In this paper, we study how to leverage answers to retrieve similar questions for queries. Specifically, we head for modeling questions and answers in a more natural way. The model should not only highlight the instinct heterogeneity of questions and answers, but also be flexible enough to take answer quality into account. To this end, we propose a supervised question-answer topic modeling approach for question retrieval in CQA. The underlying assumption is that although questions and answers are heterogeneous in many aspects, they share some common latent factors. The latent factors represent topics in askers' and answerers' minds. Following the common topics, askers ask questions in a "question language", while answerers provide answers in an "answer language". Moreover, there is a signal indicating the quality of an answer. The signal simulates how other users evaluate the answer with respect to the question in a question-answer pair. It is determined by how well askers and answerers follow common topics in the model. With the guide of quality signals, questions and answers are mapped into a common latent space (topic space) and question-question similarity can be measured with the help of information in answers compressed in that space. With the learnt topic space, we further propose a topic-based language model for question retrieval in CQA, which naturally integrates the matches in latent space and traditional retrieval models. We derive a collapsed Gibbs sampling method to estimate parameters in the model. The method is efficient in computation. Compared with existing methods, our learning approach comprehensively models user behaviors in CQA portals and at the same time naturally takes answer quality into account.

Another advantage of our method is that it is capable of leveraging rich metadata associated with answers and automatically learning answer quality signals. Specifically, we take "best answer" as the gold standard and extract a variety of features from both answer texts and associated metadata. With the training data and features, we learn a score function to generate answer quality signals and let these signals supervise the learning of the topic space. By this means, our model can not only leverage the power of big data accumulated in CQA sites, but also save on the labor of human annotators.

We test the proposed model on large scale Yahoo! Answers data and Baidu Knows data. Yahoo! Answers and Baidu Knows represent the largest and most popular CQA archives in English and Chinese, respectively. We conducted both quantitative and qualitative evaluations. The results show that our model can significantly outperform state-of-the-art translation-based approaches for question retrieval in CQA.

Our contributions in this paper are three-fold: 1) the proposal of considering answer quality when leveraging answers for question retrieval in CQA; 2) the proposal of a supervised question-answer topic model, which not only highlights the heterogeneity of questions and answers but also naturally takes answer quality into account; 3) an empirical verification of the efficacy of the proposed model on large scale English and Chinese CQA data.

The rest of the paper is organized as follows: Section 2 summarizes related work; Section 3 gives an overview of state-of-the-art methods for question retrieval in CQA; Section 4 elaborates our model, including intuition, formulation and algorithm; Section 5 reports experimental results; and finally Section 6 concludes the paper.

2. RELATED WORK

2.1 Question Retrieval in CQA

The problem of question retrieval arose from finding similar questions from frequently asked questions (FAQs). In recent years, along with the flourishing of community question answering (CQA) archives, more attention is paid to question searches in CQA. Particularly, language model based methods are proven effective. Most CQA researchers focus on leveraging metadata in CQA to improve the performance of the traditional language models for information retrieval [21]. Basically, there are two groups of work. The first group considers leveraging categories of questions. For ex-

[1]The full page is available at `https://answers.yahoo.com/question/index?qid=20080316101123AAItTLV`

ample, Cao et al. [7, 6] proposed a language model with leaf category smoothing in which they estimated a new smoothing item for language models from questions under the same category. The other group leverages answers and learns various translation models to bridge lexical gaps in questions. For example, Jeon el al. [9] proposed learning word-to-word translation probabilities from question-question pairs collected based on similar answers. Xue et al. [20] learned a translation model from question-answer pairs. In this paper, we also consider leveraging answers for retrieving similar questions in CQA. Our model is different from translation models. It unveils the common latent factors in question-answer pairs, and matches questions in the latent space. Particularly, the model takes answer quality into account, and performs robustly against noise in low quality answers.

2.2 Topic Models

We propose a topic modeling approach for question retrieval in CQA. Topic model refers to a family of probabilistic generation models that simulate the process of people writing documents and project documents into latent spaces. Early topic models like pLSI [8] assume each document is a mixture of topics and each word in a document is generated from an independent topic. Latent Dirichlet allocation (LDA) [4], which represents the state-of-the-art topic modeling approach, further assumes the distribution of topics is generated from a Dirichlet distribution. Based on LDA, many extensions have been developed for various applications. For example, Zhao and Xing [22] proposed a bilingual topic model and applied it to machine translation. Recently, supervised topic models [12, 23] were proposed to leverage side information such as labels of documents and image tags in regression or classification tasks. In this paper, we model the generation of questions and answers in a way similar to bilingual topic model. The difference is that answer quality is considered in our model.

Applying topic modeling techniques to information retrieval problems is not a new idea. Xing and Bruce [19] first applied LDA to IR and proposed an LDA-based language model for ad-hoc retrieval. In recent years, probabilistic latent aspect models have also been introduced to CQA. For example, Cai et al. [5] incorporated question category information into the traditional LDA and combined the topic model with a translation-based language model. Ji et al. [11] proposed a question-answer topic model for question retrieval in CQA. In this paper, we also attempt to apply probabilistic latent factor techniques to question retrieval in CQA. Our model is most related to the one proposed by Ji et al. The stark difference between our model and their work is that we observed the harm of low quality answers which are prevalent in popular CQA archives and incorporated answer quality signals into learning process. Therefore, our model performs more robustly against noise in answers, as will be demonstrated in Section 5.

2.3 Answer Quality in CQA

State-of-the-art question retrieval models are effective when answers have a high level of quality. Unfortunately, in practice, answer quality in popular CQA archives such as Yahoo! Answers is far from satisfying. Agichtein et al. [1] found that low quality answers take a relatively high proportion of all answers. The same phenomenon was also observed by Jeon el al. [10] and Sakai et al. [16] on popular Korean and Japanese CQA portals. To detect high quality answers, many methods have been proposed. For example, Jeon et al. [10] presented a framework to predict answer quality from non-textual features. Agichtein et al. [1] extracted both content features and usage features to learn a classifier for recognizing high quality answers. Chirag and Jefferey [17] proposed learning a logistic regression model to predict answer quality in CQA. In this paper, we employ answer quality to supervise the learning of question-answer topic space. The quality signals are automatically learnt from large scale CQA data with content and usage features used in the existing work. The automatic learning of answer quality makes our supervised topic model feasible in processing large scale CQA data.

3. PRELIMINARIES

Before presenting our model, we give a brief overview of state-of-the-art retrieval models in CQA.

3.1 Language Models for Information Retrieval

In recent years, language models for information retrieval ($LMIR$) [21] and their extensions have been proven effective for question retrieval in CQA. Formally, given a query question q and a candidate question Q, $LMIR$ calculates the similarity between q and Q by

$$P(q|Q) = \prod_{w \in q} [(1-\lambda)P_{ml}(w|Q) + \lambda P_{ml}(w|C)], \quad (1)$$

where $P_{ml}(w|Q)$ represents the maximum likelihood of term w estimated from Q, and $P_{ml}(w|C)$ is a smoothing item which is calculated as the maximum likelihood in a large corpus C. The smoothing item avoids zero probability which stems from the terms appearing in the query but not in the candidate. $\lambda \in (0,1)$ is a parameter which acts as a trade-off between the likelihood and the smoothing item.

$LMIR$ generally performs well when there is a large proportion of common terms between query q and candidate Q. In practice, however, people are used to expressing similar meanings with different words, which makes $LMIR$ suffer from the term mismatch problem when measuring similarity of questions.

3.2 Translation-based Language Model

To improve $LMIR$ when there are lexical gaps in questions, state-of-the-art methods learn word-to-word or phrase-to-phrase translation probabilities from answers, and incorporate the information into $LMIR$. Formally, given a query question q and a candidate question Q, the translation-based language model [3] is given by

$$P_{trb}(q|Q) = \prod_{w \in q} \left[(1-\lambda)P_{mx}(w|Q) + \lambda P_{ml}(w|C)\right], \quad (2)$$

where

$$P_{mx}(w|Q) = \alpha P_{ml}(w|Q) + \beta P_{tr}(w|Q)$$

$$P_{tr}(w|Q) = \sum_{v \in Q} P_{tp}(w|v) P_{ml}(v|Q).$$

Here λ, α, and β are parameters, satisfying $\alpha + \beta = 1$. $P_{tp}(w|v)$ represents the translation probability from term v in candidate question Q to term w in query question q.

Xue, et al. [20] further improved the translation-based language model by introducing an extra answer likelihood term $P_{ml}(w|a)$ estimated from the answer a of question Q. The model is defined as

$$P_{trba}(q|Q,a) = \prod_{w \in q} [(1-\lambda)P_{mx}(w|Q,a) + \lambda P_{ml}(w|C)] \quad (3)$$

where

$$P_{mx}(w|Q,a) = \alpha P_{ml}(w|Q) + \beta P_{tr}(w|Q) + \gamma P_{ml}(w|a).$$

Here γ is an extra parameter satisfying $\alpha + \beta + \gamma = 1$.

From Equation (2) and Equation (3), we can see that translation-based language models solve the term mismatch problem by linking similar words or phrases through translation probabilities. Translation probabilities are estimated from question-answer pairs under the assumption that questions and answers are parallel texts. In practice, however, questions and answers are far from "parallel". Indeed, they are highly asymmetric on the information they contain and answers are more noisy than questions. Particularly, there is usually a large amount of low quality answers in popular CQA portals. The low quality answers, like Answer 2 in Table 1, can pollute training data and make the learnt translation probabilities unreliable. Therefore, it would be better to model relationships between questions and answers in a more natural way and make the matching of questions robust to noise in answers. In the next section, we will consider mining common latent factors in questions and answers and matching questions in a latent space learnt with the help of answers and their quality signals.

4. SUPERVISED QUESTION-ANSWER TOPIC MODEL

We present matching questions in a latent space. The latent space is learnt not only from questions, but also from answers. Particularly, the learning of the latent space is guided by the quality of answers. The better the quality of an answer is, the more contributions the answer will make to the learning of the latent space. We propose a supervised question-answer topic model to learn the latent space. The model assumes that askers and answerers follow the same latent topics and write questions and answers in a "question language" and an "answer language", respectively. The latent topics further determine a signal which measures the quality of the answer with respect to the question. We interpolate the matching of questions in the latent space into the traditional $LMIR$ and propose a topic-based language model. To estimate the parameters, we derive a collapsed Gibbs sampling method which is fast in computation. Compared with translation models, our model comprehensively simulates behaviors of users on CQA archives. It not only reflects the instinct heterogeneity of questions and answers, but also highlights the importance of good answers in contrast to bad answers in learning.

4.1 A Generative Model for Questions and Answers

We propose a generative model to describe how questions and answers are created in CQA archives. Before formalizing the generative process, let us first imagine how askers, answerers, and other users interact on CQA archives. In a common case, askers have some topics on their minds and follow them to write their questions. Answerers see the questions, understand the topics of askers and follow the topics to give their answers. Finally, other users review questions and answers, and show their evaluations based on how they feel the answers fulfill needs in the questions. The evaluations usually reflect answer quality.

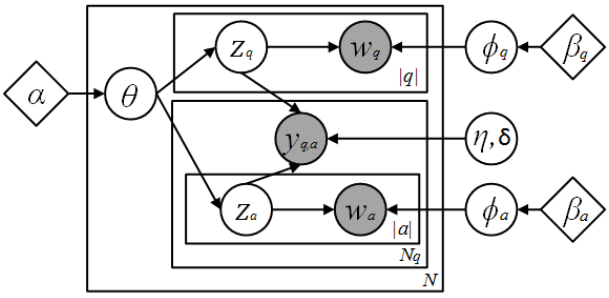


Figure 1: Supervised question-answer topic model

We try to model the interactions of users on CQA archives. Suppose that the vocabulary size is M and the total number of questions is N. Given a question q and an answer a in the answer list of q, q and a can be represented as word vectors $(w_{q,1}, \ldots, w_{q,|q|})$ and $(w_{a,1}, \ldots, w_{a,|a|})$, respectively, where $|q|$ and $|a|$ are total number of words in q and a, respectively. Suppose that the total number of answers in the answer list of q is N_q. For a fixed number of topics K, there are two $M \times K$ matrices $\phi_q = \left[\phi_q^{(1)}, \ldots \phi_q^{(K)}\right]$ and $\phi_a = \left[\phi_a^{(1)}, \ldots, \phi_a^{(K)}\right]$ representing word distributions under topics of questions and answers, respectively. In addition to these, there is a variable $y_{q,a} \in \mathbb{R}$, indicating the quality of a with respect to q. The larger $y_{q,a}$ is, the better quality a has. $y_{q,a}$ can be obtained from human labelers. In this paper, we propose automatically learning $y_{q,a}$ from CQA data, as will be seen in Section 5.1.2.

The generative process can be formalized as:

1. Draw a topic distribution $\theta \sim Dirichlet(\alpha)$
2. For word position $i = 1 : |q|$ in question q
   - Draw a topic $z_{q,i} \sim Multinomial(\theta)$
   - Draw a word $w_{q,i} \sim Multinomial(\phi_q^{(z_{q,i})})$
3. For each answer a in the answer list of q
   - For word position $j = 1 : |a|$ in answer a
     - Draw a topic $z_{a,j} \sim Multinomial(\theta)$
     - Draw a word $w_{a,j} \sim Multinomial(\phi_a^{(z_{a,j})})$
   - Draw a response variable $y_{q,a}|z_{q,1:|q|}, z_{a,1:|a|}, \eta, \delta \sim GLM(\overline{z}, \eta, \delta)$, where $\overline{z} = [1/(|q|+|a|)](\sum_{i=1}^{|q|} z_{q,i} + \sum_{j=1}^{|a|} z_{a,j})$ is the average of topic vectors, η, δ are response parameters.

Here, $GLM(\cdot,\cdot,\cdot)$ represents a generalized linear model [13]. $Dirichlet(\alpha)$ is a Dirichlet distribution with α as a hyperparameter. $\phi_q^{(z_{q_i})}$ and $\phi_a^{(z_{a_j})}$ are drawn from Dirichlet distributions $Dirichlet(\beta_q)$ and $Dirichlet(\beta_a)$, respectively, where β_q and β_a are hyperparameters. With a little manipulation of notations, we suppose that z is both a integer ranging from 1 to K and a K-dimensional indicator vector with only one element is 1 indicating the topic and all other elements are 0. The meaning of z is clear in context. The graphical model for the generative process above is given by Figure 1.

We call our model a supervised question-answer topic model. The model highlights the instinct heterogeneity of questions and answers. It reflects the asymmetry of questions

and answers in their generative processes. A question is first generated in a "question language" (i.e., ϕ_q), then answers associated with the question are generated with each in an "answer language" (i.e., ϕ_a). At the same time, the model also captures the relationships behind the heterogeneity of questions and answers, that is they are generated following the same topic distributions. Moreover, the model further considers influence of answer quality to latent topics, which simulates evaluations from other users to question-answer pairs. These characteristics differentiate our model from the existing bilingual topic model [22] and supervised topic models [12, 23].

4.2 Topic-based Language Model for Question Retrieval

With the supervised question-answer topic model, we map questions and answers into a low dimensional latent space with the supervision of answer quality signals. In the space, semantically related words in questions and answers are grouped together and noise is filtered. With the space, we can measure similarity of questions on a topic level. Formally, suppose we have obtained the posterior estimation of θ, ϕ_q, ϕ_a, given a query question q and a candidate question Q with an answer a, we interpolate matching in latent space into the traditional $LMIR$, and propose a topic-based language model as follows:

$$P_{tblm}(q|Q,a) = \prod_{w\in q}[\lambda P_{mx}(w|Q,a) + (1-\lambda)P(w|C)], \quad (4)$$

where

$$P_{mx}(w|Q,a) = \alpha P_{ml}(w|Q) + \beta P_{sqatm}(w|Q) + \gamma P_{ml}(w|a),$$
$$P_{sqatm}(w|Q) = \sum_z P(w|z,\hat{\phi})P(z|Q,\hat{\theta}).$$

Here, λ, α, β, and γ are parameters, satisfying $\lambda \in (0,1)$ and $\alpha+\beta+\gamma = 1$. $\hat{\phi}$ is defined as

$$\hat{\phi} = (1-\mu)\hat{\phi}_q + \mu\hat{\phi}_a, \quad (5)$$

where $\mu \in (0,1)$ is a combination parameter.

From Equation (4), we can see that question pairs are matched on both a term level and a topic level. The matching on the topic level is determined by a linear combination of question topic distribution ϕ_q and answer topic distribution ϕ_a, as seen in Equation (5). The impact of ϕ_q and ϕ_a in matching can be controlled through μ.

The topic-based language model is similar to the translation-based language model plus answer likelihood proposed by Xue et al. [20] (cf. Equation (3)) on appearance. The difference is that we solve the term mismatch problem through latent factors supervised by answer quality while Xue et al. rely on translation probabilities which could be unreliable when answers have low qulaity.

4.3 Parameter Estimation

We employ a Gibbs sampling approach to estimate the parameters in the supervised question-answer topic model. Specifically, we follow the work of Zhu et al. [24], introduce max-margin regularization to Bayesian inference, and develop a collapsed Gibbs sampling method. The method is efficient in computation.

Let $\boldsymbol{Z}$ denote all topics in questions and answers, $\boldsymbol{\Theta}$ denote all topic distributions, and $\boldsymbol{y}$ denote all quality signals. Define

$$\mathcal{L}(Q) = \mathrm{KL}(Q(\eta,\boldsymbol{\Theta},\boldsymbol{Z},\phi_q,\phi_a)||P_0(\eta,\boldsymbol{\Theta},\boldsymbol{Z},\phi_q,\phi_a)) - \mathbb{E}_Q[\log P(\boldsymbol{W}|\boldsymbol{Z},\phi_q,\phi_a)],$$

where $P_0(\eta,\boldsymbol{\Theta},\boldsymbol{Z},\phi_q,\phi_a)$ is a prior probability and term $\mathrm{KL}(Q(\eta,\boldsymbol{\Theta},\boldsymbol{Z},\phi_q,\phi_a)||P_0(\eta,\boldsymbol{\Theta},\boldsymbol{Z},\phi_q,\phi_a))$ is the Kullback-Leibler divergence. We try to solve the following optimization problem:

$$\arg\min_{Q\in\mathcal{P}} \mathcal{L}(Q(\eta,\boldsymbol{\Theta},\boldsymbol{Z},\phi_q,\phi_a)) + 2c\mathcal{R}_\epsilon(Q(\eta,\boldsymbol{\Theta},\boldsymbol{Z},\phi_q,\phi_a)). \quad (6)$$

Here, $\mathcal{R}_\epsilon = \sum_{i=1}^{N}\sum_{j=1}^{N_i}\mathbb{E}_Q[\max(0,|\Delta_{q_i,a_j}|-\epsilon)]$ is the expected ϵ-insensitive loss, where N is the total number of questions, for question q_i, N_i is the total number of answers in its answer list, $\Delta_{q_i,a_j} = y_{q_i,a_j} - \eta^\top \overline{z}_{q_i,a_j}$ is the margin between the quality label y_{q_i,a_j} and the prediction $\eta^\top\overline{z}_{q_i,a_j}$. $\overline{z}_{q_i,a_j}$ is the average of topic vectors in q_i and a_j. $\mathcal{P}$ is the space of probability distributions, and c is a parameter acting as a trade-off between the Bayesian inference $\mathcal{L}(Q)$ and max-margin regularization $\mathcal{R}_\epsilon$.

To solve problem (6), we define an unnormalized pseudo-likelihood of response variable:

$$\varphi(y_{q,a}|\eta,z_q,z_a) = \exp(-2c\max(0,|\Delta_{q,a}|-\epsilon))$$

Then, optimization problem (6) can be re-written as

$$\arg\min_{Q\in\mathcal{P}} \mathcal{L}\left(Q\left(\eta,\boldsymbol{\Theta},\boldsymbol{Z},\phi_q,\phi_a\right)\right) - \mathbb{E}_Q\left[\log\varphi(\boldsymbol{y}|\eta,\boldsymbol{Z})\right], \quad (7)$$

where

$$\varphi(\boldsymbol{y}|\eta,\boldsymbol{Z}) = \prod_{i=1}^{N}\prod_{j=1}^{N_i}\varphi(y_{q_i,a_j}|\eta,z_{q_i},z_{a_j}).$$

The solution of optimization problem (7) is given by

$$Q\left(\eta,\boldsymbol{\Theta},\boldsymbol{Z},\phi_q,\phi_a\right) = \frac{P_0(\eta,\boldsymbol{\Theta},\boldsymbol{Z},\phi_q,\phi_a)P(\boldsymbol{W}|\boldsymbol{Z},\phi_q,\phi_a)\varphi(\boldsymbol{y}|\eta,\boldsymbol{Z})}{\psi(\boldsymbol{y},\boldsymbol{W})}$$

where $\psi(\boldsymbol{y},\boldsymbol{W})$ is a normalizer. Following the work of Polson and Scott [14], $\varphi(y_{q,a}|\eta,z_q,z_a)$ can be represented as

$$\varphi(y_{q,a}|\eta,z_q,z_a) = \int_0^\infty \frac{1}{\sqrt{2\pi\lambda_{q,a}}}\exp\{-\frac{[\lambda_{q,a}+c(\Delta_{q,a}-\epsilon)]^2}{2\lambda_{q,a}}\}\mathrm{d}\lambda_{q,a} \times \int_0^\infty \frac{1}{\sqrt{2\pi\omega_{q,a}}}\exp\{-\frac{[\omega_{q,a}+c(\Delta_{q,a}-\epsilon)]^2}{2\omega_{q,a}}\}\mathrm{d}\omega_{q,a}$$

If we define

$$\varphi(\boldsymbol{y},\boldsymbol{\lambda},\boldsymbol{\omega}|\eta,\boldsymbol{Z}) = \prod_{i=1}^{N}\prod_{j=1}^{N_i}\frac{1}{\sqrt{2\pi\lambda_{q_i,a_j}}}\exp\{-\frac{[\lambda_{q_i,a_j}+c(\Delta_{q_i,a_j}-\epsilon)]^2}{2\lambda_{q_i,a_j}}\} \times \frac{1}{\sqrt{2\pi\omega_{q_i,a_j}}}\exp\{-\frac{[\omega_{q_i,a_j}+c(\Delta_{q_i,a_j}-\epsilon)]^2}{2\omega_{q_i,a_j}}\},$$

$Q(\eta,\boldsymbol{\Theta},\boldsymbol{Z},\phi_q,\phi_a)$ can be represented as the marginal distribution of a higher-dimensional distribution $Q(\eta,\boldsymbol{\lambda},\boldsymbol{\omega},\boldsymbol{\Theta},\boldsymbol{Z},\phi_q,\phi_a)$, which is defined as:

$$\frac{P_0(\eta,\boldsymbol{\Theta},\boldsymbol{Z},\phi_q,\phi_a)P(\boldsymbol{W}|\boldsymbol{Z},\phi_q,\phi_a)\varphi(\boldsymbol{y},\boldsymbol{\lambda},\boldsymbol{\omega}|\eta,\boldsymbol{Z})}{\psi(\boldsymbol{y},\boldsymbol{W})}.$$

Moreover, $Q(\eta,\boldsymbol{\lambda},\boldsymbol{\omega},\boldsymbol{\Theta},\boldsymbol{Z},\phi_q,\phi_a)$ is the solution of the following optimization problem:

$$\arg\min_{Q\in\mathcal{P}} \mathcal{L}\left(Q\left(\eta,\boldsymbol{\lambda},\boldsymbol{\omega},\boldsymbol{\Theta},\boldsymbol{Z},\phi_q,\phi_a\right)\right) - \mathbb{E}_Q\left[\log\varphi(\boldsymbol{y},\boldsymbol{\lambda},\boldsymbol{\omega}|\eta,\boldsymbol{Z})\right].$$

By integrating out $(\boldsymbol{\Theta},\phi_q,\phi_a)$, we have the collapsed posterior distribution $Q(\eta,\boldsymbol{\lambda},\boldsymbol{\omega},\mathbf{Z})$ as:

$$Q(\eta,\boldsymbol{\lambda},\boldsymbol{\omega},\mathbf{Z}) \propto P_0(\eta)P(\mathbf{W},\mathbf{Z}|\alpha,\beta_q,\beta_a)\varphi(\boldsymbol{y},\boldsymbol{\lambda},\boldsymbol{\omega}|\eta,\mathbf{Z}).$$

Let $n_{q,k}^{(t)}$ and $n_{a,k}^{(t)}$ denote the number of times that term t is observed with topic k in all questions and all answers, respectively; and $n_{q_i,k}$ and $n_{a_j,k}$ denote the number of times that topic k is observed in question q_i and answer a_j, respectively. Then, the conditional distributions of $Q(\eta, \boldsymbol{\lambda}, \boldsymbol{\omega}, \mathbf{Z})$ can be expressed as follows:

For η: we assume its prior is an isotropic Gaussian distribution $P_0(\eta) = \prod_{k=1}^{K} \mathcal{N}(\eta_k; 0, \nu^2)$. Then, we have

$$Q(\eta|\boldsymbol{Z}, \boldsymbol{\lambda}, \boldsymbol{\omega}) = \mathcal{N}(\eta; \mu, \Gamma) \tag{8}$$

where $\Gamma = (\frac{1}{\nu^2} I + c^2 \sum_{i=1}^{N} \sum_{j=1}^{N_i} \rho_{q_i,a_j} \overline{z}_{q_i,a_j} \overline{z}_{q_i,a_j}^{\top})^{-1}$, $\mu = c\Gamma(\sum_{i=1}^{N} \sum_{j=1}^{N_i} \psi_{q_i,a_j} \overline{z}_{q_i,a_j})$, $\rho_{q_i,a_j} = \frac{1}{\lambda_{q_i,a_j}} + \frac{1}{\omega_{q_i,a_j}}$ and $\psi_{q_i,a_j} = \frac{y_{q_i,a_j} - \epsilon}{\lambda_{q_i,a_j}} + \frac{y_{q_i,a_j} + \epsilon}{\omega_{q_i,a_j}}$.

For z: we can derive the conditional distribution of a topic z given the other topics $\boldsymbol{Z}^{\neg}$ as:

For z in question q_i:

$$\begin{aligned} &Q(z_{q_i,s}^{(k)} = 1|\boldsymbol{Z}^{\neg}, \eta, \boldsymbol{\lambda}, \boldsymbol{\omega}, w_{q_i,s} = t) \propto \\ &\frac{(n_{q,k,\neg s}^{(t)} + \beta_q)(n_{q_i,k,\neg s} + \sum_{j=1}^{N_i} n_{a_j,k} + \alpha)}{\sum_{v=1}^{M}(n_{q,k,\neg s}^{(v)} + \beta_q)} \exp\{\sum_{j=1}^{N_i} [c\gamma\psi_{q_i,a_j}\eta_k \\ &-c^2(\frac{\gamma^2 \rho_{a_i,q_j} \eta_k^2}{2} + \gamma(1-\gamma)\rho_{q_i,a_j}\eta_k \Lambda_{q_i,a_j,\neg s}^{(k)})]\}, \end{aligned} \tag{9}$$

where $w_{q_i,s}$ represents the s-th word in question q_i; $z_{q_i,s}^{(k)}$ means the topic of s-th word in question q_i is k; $n_{q,k,\neg s}^{(t)}$ is the number of times topic k is observed on term t in all questions with $w_{q_i,s}$ excluded; $n_{q_i,k,\neg s}$ is the number of times topic k is observed in question q_i with $w_{q_i,s}$ excluded. $\gamma = \frac{1}{|q_i|+|a_j|}$, and $\Lambda_{q_i,a_j,\neg s}^{(k)} = \frac{\sum_{u=1}^{K}(\eta_u(n_{q_i,u,\neg s} + n_{a_j,u}))}{|q_i|+|a_j|-1}$.

For z in answer a_j of question q_i:

$$\begin{aligned} &Q(z_{a_j,s}^{(k)} = 1|\boldsymbol{Z}^{\neg}, \eta, \boldsymbol{\lambda}, \boldsymbol{\omega}, w_{a_j,s} = t) \propto \\ &\frac{(n_{a,k,\neg s}^{(t)} + \beta_a)(n_{q_i,k} + \Omega + \alpha))}{\sum_{v=1}^{M}(n_{a,k,\neg s}^{(v)} + \beta_a)} \exp\{c\gamma\psi_{q_i,a_j}\eta_k \\ &-c^2[\frac{\gamma^2 \rho_{q_i,a_j} \eta_k^2}{2} + \gamma(1-\gamma)\rho_{q_i,a_j}\eta_k \Upsilon_{q_i,a_j,\neg s}^{(k)}]\} \end{aligned} \tag{10}$$

where $w_{a_j,s}$ represents the s-th word in answer a_j; $z_{a_j,s}^{(k)}$ means the topic of s-th word in answer a_j is k; $n_{a,k,\neg s}^{(t)}$ is the number of times topic k is observed on term t in all answers with $w_{a_j,s}$ excluded; $n_{a_j,k,\neg s}$ is the number of times topic k is observed in answer a_j with $w_{a_j,s}$ excluded. $\Omega = \sum_{l=1}^{j-1} n_{a_l,k} + n_{a_j,k,\neg s} + \sum_{l=j+1}^{N_i} n_{a_l,k}$. $\Upsilon_{q_i,a_j,\neg s}^{(k)} = \frac{\sum_{u=1}^{K}(\eta_u(n_{q_i,u} + n_{a_j,u,\neg s}))}{|q_i|+|a_j|-1}$.

For λ, ω: We can prove that λ_{q_i,a_j}^{-1} and ω_{q_i,a_j}^{-1} follow the inverse Gaussian distributions:

$$P(\lambda_{q_i,a_j}^{-1}|\boldsymbol{Z}, \eta, \boldsymbol{\omega}) = IG(\lambda_{q_i,a_j}^{-1}; \frac{1}{c|\Delta_{q_i,a_j} - \epsilon|}) \tag{11}$$

$$P(\omega_{q_i,a_j}^{-1}|\boldsymbol{Z}, \eta, \boldsymbol{\lambda}) = IG(\omega_{q_i,a_j}^{-1}; \frac{1}{c|\Delta_{q_i,a_j} - \epsilon|}) \tag{12}$$

where $IG(x; a, b) = \sqrt{\frac{b}{2\pi x^3}} \exp\{-\frac{b(x-a^2)}{2a^2 x}\}$ for $a > 0$ and $b > 0$.

With the conditional distributions above, we can construct a Markov chain to learn our model. In training, we finish the burn-in stage in T iterations based on the Markov chain, as outlined in Algorithm 1. The per-iteration time complexity of this algorithm is $O(N_{total}K + K^3)$, where N_{total} is the total number of words appearing in questions and answers.

After training, we can get the estimations of $\widehat{\theta}$, $\widehat{\phi}_q$, and $\widehat{\phi}_a$ as:

$$\begin{aligned} \widehat{\theta}_{q_i}^{(k)} &= \frac{n_{q_i,k} + \sum_{j=1}^{N_i} n_{a_j,k} + \alpha}{\sum_{u=1}^{K}(n_{q_i,u} + \sum_{j=1}^{N_i} n_{a_j,u} + \alpha)}, \\ \widehat{\phi}_{q,w}^{(k)} &= \frac{n_{q,k}^{(w)} + \beta_q}{\sum_{v=1}^{M}(n_{q,k}^{(v)} + \beta_q)}, \\ \widehat{\phi}_{a,w}^{(k)} &= \frac{n_{a,k}^{(w)} + \beta_a}{\sum_{v=1}^{M}(n_{a,k}^{(v)} + \beta_a)}, \end{aligned}$$

where $\widehat{\theta}_{q_i}^{(k)}$ is the probability of topic k in question q_i, $\widehat{\phi}_{q,w}^{(k)}$ is the probability of word w under topic k in questions, and $\widehat{\phi}_{a,w}^{(k)}$ is the probability of word w under topic k in answers.

Algorithm 1: Learning supervised question-answer topic model

```
Initialization: set all λ and ω to 1 and draw a topic z for each word in all questions and answers from a uniform distribution.
for t = 1 to T do
    draw η from the normal distribution (8)
    for each question q and its answer list do
        for each word w in question q do
            draw a topic from distribution (9)
        end for
        for each answer a in the answer list of q do
            for each word w in answer a do
                draw a topic from distribution (10)
            end for
        end for
        for each answer a in the answer list of q do
            draw λ^{-1} for a from distribution (11)
            draw ω^{-1} for a from distribution (12)
        end for
    end for
end for
```

An interesting observation is that Equation (9) reveals the effect of answer quality on the learning of question topic distribution. Specifically, the quality signal $y_{q,a}$ is encoded in $\psi_{q,a}$, which is defined as $\frac{y_{q,a} - \epsilon}{\lambda_{q,a}} + \frac{y_{q,a} + \epsilon}{\omega_{q,a}}$. The better quality an answer a has, the larger $y_{q,a}$ will be. A large $y_{q,a}$ will lead to a large $\psi_{q,a}$, which means the answer has more contributions to the learning of the question topic distribution (cf. Equation (9)).

5. EXPERIMENTS

We conducted experiments to test the performance of the proposed method on question retrieval.

5.1 Experiment Setup

5.1.1 Data Sets

We crawled a large number of questions (a.k.a. question titles) associated with descriptions (a.k.a. question bodies) and answers from Yahoo! Answers and Baidu Knows. Yahoo! Answers and Baidu Knows represent the largest and the most popular CQA archives in English and Chinese, respectively. The data covers all categories of Yahoo! Answers and Baidu Knows. For Yahoo! Answers, we crawled 1,186,542 questions with 950,127 descriptions

and 8,430,784 answers. 80% of the questions have an extra description, and on average each question is associated with 7.1 answers. For Baidu Knows, we crawled 773,194 questions with 425,257 descriptions and 2,904,596 answers. 55% of the questions have a description, and the average number of answers per question is 3.8. We used this data for training models. To prepare labeled data sets, we collected some extra questions that have been posted more recently than the training data, and randomly sampled 1,423 and 605 questions for Yahoo data and Baidu data, respectively. We took these questions as query questions. All questions, descriptions, and answers were lowercased and stemmed. Stopwords were also removed.

We separately indexed the Yahoo data and the Baidu data using an open source Lucene.Net System[2]. For each query question from Yahoo! Answers and Baidu Knows, we retrieved several candidate questions from the corresponding indexed data based on the inline ranking algorithm in Lucene.Net. On average, each query question from Yahoo! Answers has 15 candidate questions and the average number of candidate questions for Baidu data is 8.

We recruited human judges to label the relevance of the candidate questions regarding to the query questions. Specifically, for each type of language, we hired three native judges. Each judge labeled a candidate question with 1 if it is "relevant" to the query question, otherwise the judge labeled it with 0. Each candidate question got three labels and the majority of the labels was 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 ratio 1 : 3. The validation set was used for tuning parameters of different models, while the test set was used for evaluating how well the models ranked relevant candidates in contrast to irrelevant candidates.

Please note that rather than evaluate both retrieval and ranking capability of different methods like many existing works [7, 6], we compared them in a ranking task. This may lose recall for some methods, but it can enable large scale evaluation and will facilitate others reproducing our experiments and comparing their methods on our data, as we published the labeled data at `http://research.microsoft.com/en-us/people/wuwei/wuwei.aspx`.

To evaluate the performances of different models, we employed Mean Average Precision (MAP) [2], Mean Reciprocal Rank (MRR) [18], R-Precision (R-Prec) [7], and precision at position 1 (P@1) as evaluation measures. These measures are widely used in literature on question retrieval for CQA.

5.1.2 Learning Answer Quality Signals

To implement our supervised question-answer topic model, we have to collect quality signals for each answer. Ideally, these signals should be offered by human annotators. In practice, however, this would make our model infeasible on million-scale experiments. Intuitively, the performance of the proposed topic model will heavily depend on the quality of response signals. To balance the accuracy and the cost of learning, we propose automatically learning answer quality signals from CQA data.

Specifically, we extracted the following features: 1) Answer length; 2) Fraction of best answers an answerer is awarded in all answers he or she provided; 3) Unique number of words in an answer; 4) Word overlap between a question and an answer.

With these features, we took the best answer as the gold standard and trained logistic regression models [17] from the 8,430,784 answers of Yahoo! Answers and 2,904,596 answers of Baidu Knows, respectively. The advantage of logistic regression is that quality signals learnt are naturally scaled. With the automatic learning method, we not only leveraged metadata associated with answers, but also saved on the cost of human annotators.

Note that we only used four features to learn answer quality signals. The four features are effective enough for predicting answer quality. We also found that metadata like user votes is sparse and biased in Yahoo! Answers and Baidu Knows. We implemented our model with both user votes and the learnt quality signal as supervision, and compared them with other retrieval models.

For CQA sites like Quora, there are no best answers. In this case, we have to make some adjustment on the learning of answer quality signals. For example, we can take the answer with the most votes as the gold standard and train a regression model.

5.1.3 Baselines

The traditional language model for information retrieval ($LMIR$) given by Equation (1) was taken as a baseline. We also considered the following baseline methods:

Translation-based methods We learned different word-to-word translation probabilities with all the training data of Yahoo! Answers and Baidu Knows. The translation probabilities were incorporated into the model of Xue et al. given by Equation (3). Specifically, we employed the pooled approach proposed by Xue et al. [20] and estimated translation probabilities using GIZA++[3]. We took each question-answer pair as a parallel text and denoted the model as TAL_{Q-A}; we also concatenated all answers as a long document and trained a translation model from pairs of question and the long document. We denoted the model as TAL_{Q-AllA}; since best answers may have higher qulaity than other answers, we trained a model with question-best answer pairs, and denoted the model as TAL_{Q-BA}. Besides these models, we also concatenated questions and descriptions, and trained translation models using the enriched questions and answers in the same way as we did above. The three models were denoted as $TAL_{Q\&D-A}$, $TAL_{Q\&D-AllA}$, and $TAL_{Q\&D-BA}$, respectively. Finally, to make a complete comparison, we trained a model from question-description pairs, and denoted it as TAL_{Q-D}. In total, for each data set (Yahoo and Baidu data), we have 7 variants of translation models for comparison.

Topic model based methods We concatenated a question, its description and all answers associated to form a document, and ran traditional LDA [4] on the document. The learnt latent factors and topic distributions were used to calculate question similarity in a way similar to Equation (4). This method ignores the structures of questions and answers. We took it as a baseline and denoted it as $TBLM_{LDA}$. We also implemented the question-answer topic model proposed by Ji et al. [11] and considered two variants of the model: 1) the original $QATM + TransLM$,

[2] `http://lucenenet.apache.org/`

[3] http://www.statmt.org/moses/giza/GIZA++.html

Table 2: Evaluation results on Yahoo data and Baidu data

	Yahoo data				Baidu data			
	MAP	MRR	R-Prec	P@1	MAP	MRR	R-Prec	P@1
$LMIR$	0.767	0.864	0.664	0.795	0.812	0.855	0.684	0.775
TAL_{Q-D}	0.786	0.870	0.688	0.804	0.833	0.859	0.720	0.782
TAL_{Q-AllA}	0.782	0.866	0.682	0.797	0.820	0.853	0.692	0.767
TAL_{Q-A}	0.786	0.871	0.689	0.807	0.821	0.850	0.695	0.767
TAL_{Q-BA}	0.786	0.872	0.687	0.805	0.820	0.851	0.700	0.762
$TAL_{Q\&D-AllA}$	0.792	0.874	0.694	0.812	0.818	0.851	0.689	0.762
$TAL_{Q\&D-A}$	0.788	0.873	0.691	0.810	0.821	0.850	0.695	0.767
$TAL_{Q\&D-BA}$	0.791	0.875	0.694	0.813	0.819	0.856	0.686	0.769
$QATM + TransLM$	0.786	0.869	0.687	0.804	0.833	0.858	0.714	0.782
$TBLM_{LDA}$	0.788	0.875	0.690	0.811	0.838	0.868	0.715	0.793
$TBLM_{QATM}$	0.789	0.878	0.692	0.817	0.832	0.861	0.710	0.782
$TBLM_{SQATM-V}$	0.786	0.874	0.687	0.810	0.826	0.861	0.695	0.780
$TBLM_{SQATM}$	**0.805**	**0.889**	**0.718**	**0.831**	**0.869**	**0.894**	**0.771**	**0.834**

which is the best performing model in [11][4]; 2) we calculated $P_{sqatm}(w|Q)$ in Equation (4) with the latent factors and topic distributions learnt by QATM, and denoted the model as $TBLM_{QATM}$. Parameters in all topic models were estimated with Gibbs sampling approaches.

We denoted our model using the learnt quality signal as $TBLM_{SQATM}$. We also implemented the model with user votes as supervision, and denoted it as $TBLM_{SQATM-V}$. We compared our model with all baseline methods on the test data sets.

5.1.4 Parameter Tuning

There are several parameters we have to determine in our experiments. For $LMIR$ and translation-based methods, we tuned the smoothing parameters $\{\lambda, \alpha, \beta, \gamma\}$ in $\{0.1, 0.2, \cdots, 0.9\}$. For topic model based methods, following [24], we used the symmetric Dirichlet priors $\boldsymbol{\alpha} = \frac{1}{K}\mathbf{1}$, $\boldsymbol{\beta_q} = \boldsymbol{\beta_a} = 0.01 \times \mathbf{1}$, where $\mathbf{1}$ is a vector with all entries 1. We tuned the number of topics (i.e.,K) in {10, 20, 30, 40, 50} and the number of iterations of Gibbs sampling (i.e., T) in $\{100, 200, \cdots, 1000\}$. The best parameter combinations for LDA and QATM are (30, 1000) and (50, 1000) respectively on both data sets. After we incorporated LDA and QATM into Equation (4), the smoothing parameters $\{\lambda, \alpha, \beta, \gamma\}$ also need tuning. We tuned them in the same way as in translation-based methods. In $QATM + TransLM$, there is an extra parameter μ. We tuned it in $\{0.1, 0.2, \cdots, 0.9\}$.

In our supervised question-answer topic model, besides number of topics and number of iterations, we used the standard normal prior with $\nu^2 = 1$ in $P_0(\eta)$ and set $\epsilon = 1e^{-3}$ in the expected ϵ-insensitive loss $\mathcal{R}_\epsilon$. c in optimization problem (6) is selected from {0.001, 0.01, 0.1, 1, 10, 100, 1000}. μ in Equation (5) is tuned in {0, 0.1, 0.2, ..., 0.9, 1}. The best parameter combination for $TBLA_{SQATM}$ is $(K = 40, T = 1000, c = 1, \mu = 0.8)$ and for $TBLA_{SQATM-V}$ is $(K = 40, T = 1000, c = 1, \mu = 0.7)$.

5.2 Quantitative Evaluation

Table 2 shows the quantitative evaluation results of different models on Yahoo data and Baidu data, respectively. From Table 2, we can see that $TBLM_{SQATM}$ outperforms all baseline methods on all evaluation metrics. When we conducted a statistical test (t-test), the results show that the improvements are statistically significant (p value < 0.01).

$TBLM_{SQATM}$ significantly outperforms $LMIR$, which demonstrates that matching questions with latent topics estimated from question-answer pairs can effectively address the term mismatch problem. Its significant improvements over translation-based methods and other topic model based methods further verify that incorporating answer quality into learning can calibrate the learnt latent space and make the model robust to noise in answers.

$TBLM_{SQATM-V}$ performs badly on both data sets. We analyzed the reason and found that user votes are very sparse and biased on Yahoo and Baidu data. Many answers with high quality only received a few votes, while some bad answers, due to their funny words, attracted much attention. Moreover, most answers have 0 votes, because they are posted late. Therefore, in many cases, $TBLM_{SQATM-V}$ degenerated to $TBLM_{QATM}$, and sometimes it is even worse, due to unreliable quality supervision.

On both data sets, topic model based methods perform better than most translation-based methods. The results demonstrate that the latent factor approach can model CQA data, especially relationships between questions and answers, in a better way than simply taking questions and answers as parallel pairs, and matching questions via latent factors can be a better alternative for bridging lexical gaps in questions than matching questions through translation probabilities estimated from question-answer pairs.

Translation-based methods perform inconsistently on Yahoo data and Baidu data. On Yahoo data, the best performing model is $TAL_{Q\&D-AllA}$, while on Baidu data, the best model is TAL_{Q-D}. This result may stem from the different average numbers of answers per question in the two data sets (7.1 in Yahoo data v.s. 3.8 in Baidu data). When there are only a few answers, the influence of noise in answers is enlarged and the translation from answers becomes more unreliable. Learning only from best answers can be a way to alleviate the hurt of low quality answers, however, this method will also lose some useful information. Therefore, on Baidu data, translation with question-description pairs performs even better than translation with question-answer pairs. The same phenomenon is observed on the comparison of $TBLM_{LDA}$ and $TBLM_{QATM}$. In fact, when answers are insufficient, the question-answer topic model is more vulnerable to low quality answers than the traditional LDA. This

[4]We did not implement QATM with posterior regularization, as its performance is comparable with QATM in [11].

further emphasized the importance of considering answer quality in learning.

5.3 Qualitative Evaluation

We investigate why $TBLM_{SQATM}$ can outperform baseline methods on question retrieval.

Table 3(a) gives an example for comparing $TBLM_{SQATM}$ and $LMIR$. Candidate question "What is the eye dr. looking for when he/she dilates my eyes?" is relevant to query question "what does a doctor look for when they dilate your eyes", but they share a relatively small proportion of common words. Particularly, "doctor" in query mismatches with "dr." in candidate which is a abbreviation of "doctor". On the other hand, another irrelevant candidate "When doctors dilate your eyes, what do they use?" shares a large proportion of common words with the query question. $LMIR$ measures question similarity only based on the common terms they share. Therefore, it ranked the irrelevant one at a higher position than the relevant one. Our method $TBLM_{SQATM}$, on the other hand, can group semantically related words like "doctor", "dr.", "nurse","medical", and "cancer", et al. together in a latent topic space, and successfully recognize the similarity between the relevant candidate and the query.

Table 3(b) gives an example for comparing $TBLM_{SQATM}$ and $TAL_{Q\&D-AllA}$(performed best on Yahoo data). Similar to Table 3(a), query question and relevant candidate question express similar meanings with different words. In the query, "infants" was used, while in the relevant candidate the same information need was expressed by the word "baby". Both $TBLM_{SQATM}$ and $TAL_{Q\&D-AllA}$ attempted to leverage answers to bridge the lexical gaps. However, we found that translation probabilities were polluted by noise from answers. In the translation table of $TAL_{Q\&D-AllA}$, "infant" can only be translated from words like "old", "carrier", "hear", "live-in", "opinion", and "see" with high probabilities. None of these words appear in the relevant candidate, and $TAL_{Q\&D-AllA}$ degenerated to $LMIR$. Therefore, $TAL_{Q\&D-AllA}$ ranked the irrelevant question which shares more common terms with the query at a higher position. On the other hand, for $TBLM_{SQATM}$, we checked the topic distributions estimated for the relevant candidate, and found that in the space of the most significant topic, semantically related words like "baby", "body", "child", "women", "birth" and "infant" were grouped together and got high generative probabilities. Therefore, although "infants" and "baby" mismatch with each other, their semantic relationship was captured by $TBLM_{SQATM}$. With answer quality signals, $TBLM_{SQATM}$ performed more robustly against the noise in answers, and was capable of leveraging useful information in good answers to solve the term mismatch problem.

Table 3(c) gives an example for comparing $TBLM_{SQATM}$ and $TBLM_{QATM}$. In this case, the relevant candidate is long and contains some related sense like "burned movie" and "dvd rom". Therefore, on the term level, the short irrelevant question seems more likely a good candidate. To capture the semantic similarity between the candidates and the query question, both $TBLM_{SQATM}$ and $TBLM_{QATM}$ try to leverage the topic space. However, when we checked the topic distributions estimated for the relevant candidate, we found that the most significant topic of $SQATM$ contains words like "pc", "computer", "cd", "play", "watch", "phone", "ipod", "recognize", "window" and "dvd" with high probabilities, while in the most significant topic of $QATM$, words like "game", "car", "yahoo", "home" "pc", "computer", "box", "friend", "run" and "dvd" dominated. It is clear that the topic space of $QATM$ was polluted by the noise in low quality answers. With answer quality signals, $SQATM$ can learn a better topic representation than $QATM$, and successfully recognize the similarity between the relevant candidate and the query question.

Table 3: Examples for comparison between $TBLM_{SQATM}$ and baselines

(a) Compare $TBLM_{SQATM}$ with $LMIR$

Query	what does a ***doctor*** look for when they dilate your eyes	
Label	Relevant	Irrelevant
Question	What is the eye ***dr.*** looking for when he/she dilates my eyes?	When ***doctors*** dilate your eyes, what do they use?
$LMIR$	Rank3	Rank2
$TBLM_{SQATM}$	Rank2	Rank3

(b) Compare $TBLM_{SQATM}$ with $TAL_{Q\&D-AllA}$

Query	how to get ***infants*** to take medicine	
Label	Relevant	Irrelevant
Question	How to get a ***baby*** to take medicine?	Did they take all ***infant*** cold medicines off the shelf, or not?
$TAL_{Q\&D-AllA}$	Rank3	Rank2
$TBLM_{SQATM}$	Rank2	Rank3

(c) Compare $TBLM_{SQATM}$ with $TBLM_{QATM}$

Query	Why wont my computer recognize my dvd?	
Label	Relevant	Irrelevant
Question	Why can i only watch burned movies on my dvd rom. it wont recognize a regular dvd??	My computer wont recognize Fallout 3 When i put disk in?
$TBLM_{QATM}$	Rank3	Rank2
$TBLM_{SQATM}$	Rank2	Rank3

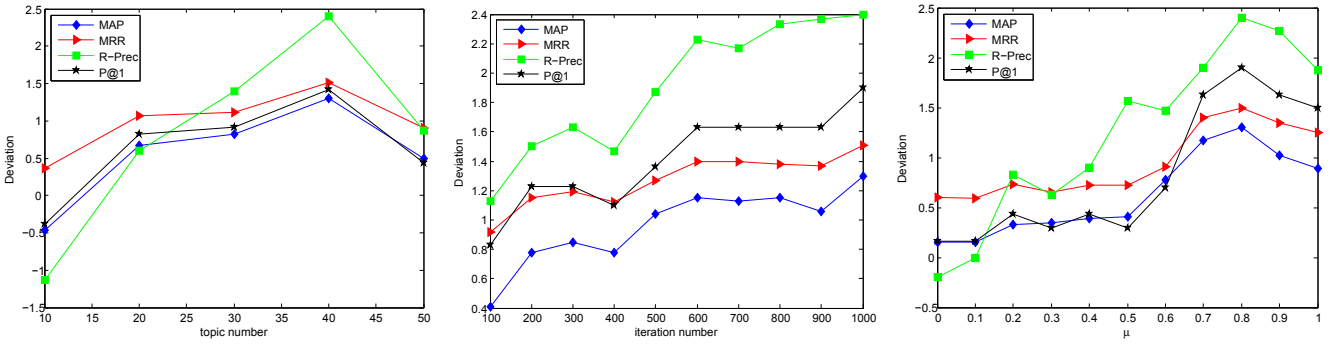
5.4 Discussion

We studied how the number of topics (i.e., K), the number of iterations in training (i.e., T), and the trade-off between question topic and answer topic (i.e., μ in Equation (5)) influence the performance of $TBLM_{SQATM}$. All analysis was done on Yahoo data, since it is larger than Baidu data.

For topic number K, intuitively, too few topics are not sufficient to capture the latent structures in questions and answers, while too many topics will introduce noise to the learnt latent space. We tried different numbers of K. Figure 2(a) gives the results. The results verified our intuitions and indicated that 40 is the best choice of topic number.

For iteration number T, on one hand, it is preferred to iterate until convergence; on the other hand, early stopping, as a kind of regularization, could lead to better performance of the model. We tried different numbers of iterations on $TBLM_{SQATM}$. Figure 2(b) shows the result. From the figure, we did not see evident influence of early stopping and the sampling method quickly get converged after 600 iterations.

Finally, Figure 2(c) illustrates the influence of μ to the performance of $TBLM_{SQATM}$. From the figure, we can see that a large μ will help improve the relevance and $\mu = 0.8$



(a) Influence of topic numbers (b) Influence of iteration numbers (c) Influence of μ

Figure 2: The influences of different factors on the performance of $TBLM_{SQATM}$ on Yahoo data. y axis represents the difference of $TBLM_{SQATM}$ and $TAL_{Q\&D-AllA}$ on evaluation metrics.

is the best. Thus, we can conclude that answer topics contribute more in solving the term mismatch problem. This result is reasonable, because there is richer information in answers than in questions, and our model can effectively mine the useful information from answers and leverage the information to match questions in a better way.

6. CONCLUSION

We propose a supervised question-answer topic model for question retrieval in CQA. With the model, we match questions in a topic space learned from question-answer pairs under the guide of answer quality signals. Experiments on million scale real world CQA data verify the efficacy of the proposed model.

7. ACKNOWLEDGMENT

This work was supported by NSFC (Grand Nos. 61170189, 61370126, 61202239), the Research Fund for the Doctoral Program of Higher Education (Grand No. 20111102130003), the Fund of the State Key Laboratory of Software Development Environment (Grand No. SKLSDE-2013ZX-19), and Microsoft Research Asia Fund (Grand No. FY14-RES-OPP-105).

8. REFERENCES

[1] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media. In *WSDM'08*, pages 183–194, 2008.

[2] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999.

[3] A. Berger, R. Caruana, D. Cohn, D. Freitag, and V. Mittal. Bridging the lexical chasm: statistical approaches to answer-finding. In *SIGIR'00*, pages 192–199, 2000.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *JMLR'03*, 3:993–1022, 2003.

[5] L. Cai, G. Zhou, K. Liu, and J. Zhao. Learning the latent topics for question retrieval in community qa. In *IJCNLP'11*, pages 273–281, 2011.

[6] X. Cao, G. Cong, B. Cui, and C. S. Jensen. A generalized framework of exploring category information for question retrieval in community question answer archives. In *WWW'10*, pages 201–210, 2010.

[7] X. Cao, G. Cong, B. Cui, C. S. Jensen, and C. Zhang. The use of categorization information in language models for question retrieval. In *CIKM'09*, pages 265–274, 2009.

[8] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR'99*, pages 50–57, 1999.

[9] J. Jeon, W. B. Croft, and J. H. Lee. Finding similar questions in large question and answer archives. In *CIKM'05*, pages 84–90, 2005.

[10] J. Jeon, W. B. Croft, J. H. Lee, and S. Park. A framework to predict the quality of answers with non-textual features. In *SIGIR'06*, pages 228–235, 2006.

[11] Z. Ji, F. Xu, B. Wang, and B. He. Question-answer topic model for question retrieval in community question answering. In *CIKM'12*, pages 2471–2474, 2012.

[12] J. D. Mcauliffe and D. M. Blei. Supervised topic models. In *NIPS'07*, pages 121–128, 2007.

[13] P. McCullagh. Generalized linear models. *European Journal of Operational Research*, 16(3):285–292, 1984.

[14] N. G. Polson and S. L. Scott. Data augmentation for support vector machines. *Bayesion Analysis*, 6(1):1–24, 2011.

[15] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at trec-3. In *TREC*, 1994.

[16] T. Sakai, D. Ishikawa, N. Kando, Y. Seki, K. Kuriyama, and C.-Y. Lin. Using graded-relevance metrics for evaluating community qa answer selection. In *WSDM'11*, pages 187–196, 2011.

[17] C. Shah and J. Pomerantz. Evaluating and predicting answer quality in community qa. In *SIGIR'10*, pages 411–418, 2010.

[18] E. M. Voorhees. The trec-8 question answering track report. In *Proceedings of the 8th Text Retrieval Conference*, pages 77–82, 1999.

[19] X. Wei and W. B. Croft. Lda-based document models for ad-hoc retrieval. In *SIGIR'06*, pages 178–185, 2006.

[20] X. Xue, J. Jeon, and W. B. Croft. Retrieval models for question and answer archives. In *SIGIR'08*, pages 475–482, 2008.

[21] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, 2004.

[22] B. Zhao and E. P. Xing. Bitam: Bilingual topic admixture models for word alignment. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 969–976, 2006.

[23] J. Zhu, A. Ahmed, and E. P. Xing. Medlda: maximum margin supervised topic models for regression and classification. In *ICML'09*, pages 1257–1264, 2009.

[24] J. Zhu, N. Chen, H. Perkins, and B. Zhang. Gibbs max-margin topic models with fast sampling algorithms. In *ICML'13*, pages 124–132, 2013.