Learning to Suggest Questions in Online Forums

Tom Chao Zhou[1*], **Chin-Yew Lin**[2], **Irwin King**[3†],
Michael R. Lyu[1], **Young-In Song**[2] and **Yunbo Cao**[2]
[1]Department of Computer Science and Engineering,
The Chinese University of Hong Kong, Shatin, N.T., Hong Kong
[2]Microsoft Research Asia, Beijing, China
[3]AT&T Labs Research, Florham Park, NJ 07932, USA
[1]{czhou, lyu}@cse.cuhk.edu.hk [3]irwin@research.att.com
[2]{cyl, yosong, yunbo.cao}@microsoft.com

Abstract

Online forums contain interactive and semantically related discussions on various questions. Extracted question-answer archive is invaluable knowledge, which can be used to improve Question Answering services. In this paper, we address the problem of Question Suggestion, which targets at suggesting questions that are semantically related to a queried question. Existing bag-of-words approaches suffer from the shortcoming that they could not bridge the lexical chasm between semantically related questions. Therefore, we present a new framework to suggest questions, and propose the Topic-enhanced Translation-based Language Model (TopicTRLM) which fuses both the lexical and latent semantic knowledge. Extensive experiments have been conducted with a large real world data set. Experimental results indicate our approach is very effective and outperforms other popular methods in several metrics.

1 Introduction

An online forum is a Web application which involves highly interactive and semantically related discussions on domain specific questions, such as travel, sports, programming. Questions are usually the focus of forum discussions and a natural means of resolving issues (Shrestha and McKeown 2004). Previous research efforts show that mining forum knowledge in the form of Question-Answer (QA) pairs could improve forum management(Cong et al. 2008). Over times, a large amount of historical QA pairs have been built up in forum archives, providing information seekers a viable alternative to general purpose Web search (Bian et al. 2008).

A user posts or searches a query in forum archives because he/she is interested in a particular topic, while unaware that his/her query may only capture one aspect of the particular topic. However, existing services only provide "question search", which targets at finding semantically equivalent questions to a query. An example of question search is shown in Table 1. In Table 1, the user's query is *How is Orange Beach in Alabama?*. He/she may not be aware that an existing question *Is the water pretty clear this time of year on Orange Beach?* can also satisfy his/her information needs. Under these circumstances, it is necessary and desirable to suggest semantically related questions. Good question suggestion has three benefits: (1) helping users explore their information needs thoroughly from different aspects; (2) increasing page views by enticing users' clicks on suggested questions; (3) providing forums a relevance feedback mechanism by mining users' click through logs. Existing methods in question search only employ bag-of-words approach with lexical knowledge, failing to bridge the lexical chasm between semantically related questions (Jeon, Croft, and Lee 2005) (Xue, Jeon, and Croft 2008). This paper proposes an effective question suggestion framework in online forums. The framework is shown in Fig. 1. Specifically, this framework consists of three major steps: (1) detecting questions in forum threads; (2) learning word translation probabilities from questions in forum threads; (3) calculating semantic relatedness between a queried question and a candidate question using Topic-enhanced Translation-based Language Model (TopicTRLM).

Table 1: **Question Search and Suggestion Examples.**

Query: How is Orange Beach in Alabama?
Question Search: Any ideas about Orange Beach in Alabama?
Question Suggestion: Is the water pretty clear this time of year on Orange Beach? Do they have chair and umbrella rentals on Orange Beach?

The rest of the paper is organized as follows. Section 2 introduces the related work. In Section 3, we present the proposed approach to question suggestion. In Section 4, we empirically verify the effectiveness of the proposed approach. Section 5 summarizes our work and discusses future work.

2 Related Work

Our work is related to question search. Translation model has been extensively employed in question search (Jeon, Croft, and Lee 2005) (Duan et al. 2008). Realizing that translation model may produce inconsistent probability estimates and make the model unstable, Xue et al. (Xue, Jeon, and Croft 2008) proposed translation-based language model

[*] This work was done when Tom Chao Zhou was an intern at Microsoft Research Asia.

[†] Currently, Irwin King is on leave from The Chinese University of Hong Kong.

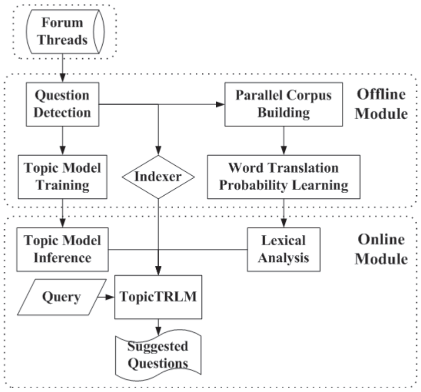


Figure 1: **System framework of question suggestion.**

which balances between language model and translation model. Besides using translation model, Cao et al. (Cao et al. 2010) proposed to use categorization information in question retrieval task, and Wang et al. (Wang, Ming, and Chua 2009) proposed a syntactic tree matching approach to find similar questions. Cao et al. (Cao et al. 2008) proposed the MDL-based tree cut model for question recommendation. Our work has two differences comparing with previous approaches. Firstly, the proposed TopicTRLM fuses both the lexical and latent semantic information to improve question suggestion; while previous methods only employed lexical knowledge. Secondly, our work proposes an effective method to build a parallel corpus of related questions by utilizing the interactive nature of online forums.

3 Question Suggestion

Questions are usually the focus of forum discussions and a natural means of resolving issues. In this paper, we adopt the method used in (Cong et al. 2008) for question detection.

Two types of methods are typically used to represent the content of text documents. One is the bag-of-words representation, which means that words are assumed to occur independently. A bag-of-words model is a fine-grained representation of a text document. The other method to represent text documents is topic model. Topic model assigns a set of latent topic distributions to each word by capturing important relationships between words. Comparing with bag-of-words representation, topic model is a coarse-grained representation for documents.

Suggested questions should be semantically related to the queried question, and they should explore different aspects of a discussion topic with respect to the queried question. Fine-grained bag-of-words representation of question would contribute to finding lexically similar questions, and topic model representation would contribute to finding semantically related questions. To achieve the goal of adopting both bag-of-words and topic model representations, we propose the TopicTRLM model. It fuses the latent topic information with lexical information to measure the semantic relatedness between two questions systematically. Specifically, we employ the Translation-based Language Model (TRLM) to measure the semantic relatedness of bag-of-words representations of two questions and employ Latent Dirichlet Allocation (LDA) to calculate the latent topics' similarities between two questions.

Equation (1) shows TopicTRLM approach to calculate the semantic relatedness of a queried question and a candidate question:

$$\begin{aligned} P(q|D) &= \prod_{w \in q} P(w|D), \\ P(w|D) &= \gamma P_{trlm}(w|D) + (1-\gamma) P_{lda}(w|D), \quad (1) \end{aligned}$$

where q is the queried question, D is a candidate question, w is a query term in q, $P_{trlm}(w|D)$ is the TRLM score, and $P_{lda}(w|D)$ is the LDA score. Equation (1) employs Jelinek-Mercer smoothing (Zhai and Lafferty 2004) to fuse the TRLM score with LDA score, and γ is the parameter to balance the weights of bag-of-words representation and topic-model representation. A larger γ means that we would like to find more lexically related questions for the queried question; a smaller γ would emphasize more on two questions' latent topic distributions' similarity. When we set $\gamma = 0$, TopicTRLM only employs latent topic analysis, and when we set $\gamma = 1$, TopicTRLM only employs lexical analysis. Thus, TopicTRLM is a generalization of both lexical analysis and latent topic analysis in the question suggestion task. Equation (2) describes TRLM which employs Dirichlet smoothing:

$$\begin{aligned} P_{trlm}(w|D) &= \frac{|D|}{|D|+\lambda} P_{mx}(w|D) + \frac{\lambda}{|D|+\lambda} P_{mle}(w|C), \\ P_{mx}(w|D) &= \beta P_{mle}(w|D) + \\ &\quad (1-\beta) \sum_{t \in D} T(w|t) P_{mle}(t|D), \quad (2) \end{aligned}$$

where $|D|$ is the length of the candidate question, C is the question collection extracted from the forum posts. λ is the Dirichlet smoothing parameter to balance the collection smoothing and empirical data. If we increase the λ, then we would rely more on smoothing. Dirichlet smoothing has the advantage that for longer candidate questions. Its smoothing effect would be smaller. β is the parameter to balance between language model and translation model. A larger β would have the effect to retrieve lexically similar questions. A smaller β would have the effect to retrieve lexically related questions. $T(w|t)$ is the translation probability from source word t to target word w, $P_{mle}(\cdot)$ is the maximum likelihood estimation. An essential part of TRLM is to learn the word to word translation probabilities $T(w|t)$, which would be discussed later. Equation (3) describes employing LDA to calculate the similarity between a query term w and a candidate D:

$$P_{lda}(w|D) = \sum_{z=1}^{K} P(w|z) P(z|D), \quad (3)$$

where K is the number of latent topics, and z is a latent topic.

Learning Translation Probabilities in Forums: Learning word to word translation probabilities is the most essential part to employ TRLM. IBM model 1 (Brown et al.

1990) is employed to learn the translation probabilities, and a monolingual parallel corpus is needed. The construction of the parallel corpus should be tailored to the specific task. To find similar questions, three kinds of approaches are employed previously to build parallel corpus: (1) question and question pairs are considered as a parallel corpus if their answers are similar (Jeon, Croft, and Lee 2005), (2) question and answer pairs are considered as a parallel corpus (Xue, Jeon, and Croft 2008), and (3) question and its manually labeled question reformulation pairs are considered as a parallel corpus (Bernhard and Gurevych 2009). However, neither of above three methods is suitable to build the parallel corpus for the question suggestion task in forums. The reason is that the presence of spam within the discussion forum would make all questions subjected to the same spam appear equivalent. To build a parallel corpus for learning word to word translation probabilities for question suggestion, we turn to investigating the properties of forum discussions. Because questions are usually the focus of forum discussions and a natural means of resolving issues, questions posted by a thread starter during the discussion are very likely to explore different aspects of a topic. It is very likely that these questions are semantically related. Thus, we propose to utilize these semantically related questions posted by the thread starter in each thread to build the parallel corpus. The procedure of generating a parallel corpus of related questions from forums is as follows: (1) extract questions posted by the thread starter in a thread, and create a question pool Q; (2) construct question-question pairs by enumerating all possible combinations of question pairs in the Q; (3) repeat step 1 and 2 for each forum thread; (4) build the parallel corpus by aggregating all question-question pairs constructed from each forum thread.

Latent Dirichlet Allocation: Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003), as a topic model method that possesses fully generative semantics, has attracted a lot of interests in the machine learning field. The graphical model of LDA is shown in Fig. 2. The process of generating a corpus in the smoothed LDA is as follows: (1) pick a multinomial distribution ϕ_z for each topic z from a Dirichlet distribution with parameter β; (2) pick a multinomial distribution θ_D from a Dirichlet distribution with parameter α for each question D; (3) pick a topic $z \in \{1, \ldots, K\}$ from the multinomial distribution θ_D for each word token w in question D; (4) pick word w from the multinomial distribution ϕ_z.

We calculate the semantic relatedness between a query word w and a candidate question D as follows:

$$p_{lda}(w|D, \hat{\theta}, \hat{\phi}) = \sum_{z=1}^{K} p(w|z, \hat{\phi}) p(z|\hat{\theta}, D), \tag{4}$$

where $\hat{\theta}$ and $\hat{\phi}$ are the posterior estimates of θ and ϕ. We employ Gibbs sampling (Griffiths and Steyvers 2004) to directly obtain the approximation of $\hat{\theta}$ and $\hat{\phi}$ because the LDA model is quite complex and cannot be solved by exact inference. In a Gibbs sample, $\hat{\phi}$ is approximated with $(n_{-i,j}^{(w_i)} + \beta_{w_i}) / \sum_{v=1}^{V} (n_{-i,j}^{(v)} + \beta_v)$, and $\hat{\theta}$ is approximated with $(n_{-i,j}^{(D_i)} + \alpha_{z_i}) / \sum_{m=1}^{M} (n_{-i,m}^{(D_i)} + \alpha_m)$ after a certain number of iterations being accomplished. $n_{-i,j}^{(w_i)}$ is the number of instances of word w_i assigned to topic $z = j$, not including the current token. α and β are hyper-parameters that determine how heavily this empirical distribution is smoothed. $n_{-i,j}^{(D_i)}$ is the number of words in document D_i assigned to topic $z = j$, not including the current token. The total number of words assigned to topic $z = j$ is $\sum_{v=1}^{V} n_{-i,j}^{(v)}$. The total number of words in document D not including the current one is $\sum_{m=1}^{M} n_{-i,m}^{(D_i)}$. Based on these derivations, we rewrite Eq. (3) as Eq. (5):

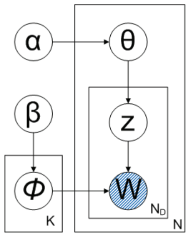


Figure 2: **Graphical model of LDA. N is the number of documents; N_D is the number of words in document D; K is the number of topics.**

$$P_{lda}(w|D) = \sum_{z=1}^{K} \frac{n_{-i,j}^{w_i} + \beta_{w_i}}{\sum\limits_{v=1}^{V} (n_{-i,j}^{(v)} + \beta_v)} \times \frac{n_{-i,j}^{D_i} + \alpha_{z_i}}{\sum\limits_{m=1}^{M} (n_{-i,m}^{(D_i)} + \alpha_m)}. \tag{5}$$

4 Experiments and Results

We consider the question suggestion task as a retrieval task in our experiments. We aim to address three research questions through our experiments:

RQ1: How effective is the proposed method to learn the word to word translation probabilities?

RQ2: How is TopicTRLM compared with other approaches on labeled questions in question suggestion task?

RQ3: How is TopicTRLM compared with other approaches on the joint probability distributions' similarity of topics with ground truth?

4.1 Experimental Setup

Methods: To evaluate the performance of the proposed methods, we compared the proposed algorithms with alternative approaches. Specifically, we compared LDA (Blei, Ng, and Jordan 2003), query likelihood language model using Dirichlet smoothing (QL) (Zhai and Lafferty 2004), translation model (TR) (Jeon, Croft, and Lee 2005), and the state-of-the-art question search method translation-based language model (TRLM) (Xue, Jeon, and Croft 2008).

Data set: For evaluation purpose, we crawled data from the travel forum TripAdvisor[1]. TripAdvisor is a popular online forum that attracts a large number of discussions about

[1] http://www.tripadvisor.com

hotels, traveler guides, etc. TripAdvisor forum consists of a large number of threads, which contain posts from thread starters and other participants. The crawling process was conducted from the thread level. We employed the same settings with (Cong et al. 2008) to mine LSPs, and the classification-based question detection method was reported to score 97.8% in Precision, 97.0% in Recall, and 97.4% in F_1-score.

After employing the question detection method in crawled data, we randomly sampled 300 questions, we removed questions that are not comprehensible, e.g., *What to see?* is not a comprehensible question; while *How is the Orange Beach in Alabama?* is a comprehensible question. Finally we got 268 questions. We used the unigram language model to represent questions, and applied IBM model 1 to learn unigram to unigram translation probabilities. We used Porter Stemmer to stem question words. We adopted the stop word list used by SMART system, but 5W1H words were removed from the stop word list. For each model, the top 20 retrieval results were kept. We used *pooling* to put results from different models for one query together for annotation, and all models were used in the pooling process. If a returned result was considered as semantically related to the queried question, it was labeled with "relevant"; otherwise, it was labeled with "irrelevant". Two assessors were involved in the initial labeling process. If two assessors had different opinions on a decision, a third assessor was asked to make a final decision. The kappa statistics between two assessors was 0.74. This test set was referred to as "TST_LABEL".

In order to create a reasonable ground truth data without involving laborious manual labeling, we assumed that questions posted by the same user in a thread were related. We built the unlabeled testing data set by randomly selecting threads until there were $10,000$ threads that contain at least two questions posted by thread starters. The first question in each thread was treated as the queried question. This test set was referred to as "TST_UNLABEL".

The remaining questions, referred to as "TRAIN_SET", were used in three purposes: (1) building parallel corpus to learn the word to word translation probabilities, (2) LDA training data, and (3) question repository to retrieve questions to offer question suggestion service. TRAIN_SET contained $1,976,522$ questions extracted from $971,859$ threads. We conducted a detailed analysis on the TRAIN_SET to get a deeper understanding of the forum activities.

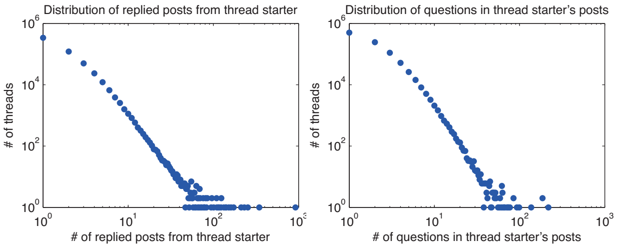
This paper leveraged thread starters' activities in forums, so we first conducted a post level analysis on thread starters' activities. The statistics is shown in Table 2. From Table 2, we can see thread starters replied on average 1.9 posts to the thread he or she initiated, and this indicates our expectation that forum discussions are quite interactive. We also plotted the distribution of replied posts from thread starter in Fig. 3(a), and this distribution follows a power law distribution. In addition, this is the first time the power distribution of thread starters' activities is reported. We also conducted a question level analysis on thread starters' activities. Table 3 presents statistics of question level activities of thread starter. We found over 68.8% thread starters asked on average 2 questions in each thread. These findings supported our motivation that question is a focus of forum discussions, and forum data is an ideal source to train the proposed model for question suggestion. Figure 3(b) depicts a view of distribution of questions in thread starter's posts.

Table 2: **Statistics of Post Level Activities of Thread Starter (TS).**

#Threads	#Threads that have replied posts from TS	Avg.# replied posts from TS
1,412,141	566,256	1.9

Table 3: **Statistics of Question Level Activities of Thread Starter (TS).**

#Threads	#Threads TSs' posts contain questions	Avg.# questions in TSs' posts
1,412,141	971,859	2.0



(a) **Post level.** (b) **Question level.**

Figure 3: **Distribution of Thread Starters' Activity.**

We used GIZA++ (Och and Ney 2003)[2] to train the IBM model 1. We used GibbsLDA++ (Phan, Nguyen, and Horiguchi 2008)[3] to conduct LDA training and inference.

Metrics: For the evaluation of the task, we adopted several well known metrics that evaluate different aspects of the performance of the proposed method, including Precision at Rank R ($P@R$), Mean Average Precision (MAP), Mean Reciprocal Rank (MRR), and Kullback-Leibler divergence (KL-divergence).

Parameter Tuning: There are several parameters need to be determined in our experiments. We used 20 queries from the TST_LABEL, and employed MAP to tune the parameters. Optimal parameters are as follows: $\alpha = \frac{50}{K}$, $\beta = 0.1$, $K = 200$, $\lambda = 2,000$, $\beta = 0.2$ and $\gamma = 0.7$.

4.2 Experiment on Word Translation

To answer RQ1, we used the proposed method to build the parallel corpus, and the constructed parallel corpus contains $2,629,533$ question-question pairs. Table 4 shows the top 10 words that are most semantically related to the given words employing IBM model 1 and LDA.

Various semantic relationships between words were discovered using IBM model 1. For example, when a user is asking a question about *shore*, *snorkel* is related because *snorkelling* is a popular activity in *shore*, and *condo* is also related because the user also needs to rent a condo for liv-

[2] http://fjoch.com/GIZA++.html

[3] http://gibbslda.sourceforge.net

ing. *Walton* is a beach name in Florida's Emerald Coast near *Pensacola* and *Destin*. Its full name is *Fort Walton Beach*. *Atlanta* is also related to *Walton* because the nearest Airport of *Walton* provides frequent flights to *Atlanta*. Recall that the proposed method considers that questions in a thread could translate to each other, leading to capturing the semantic relationships of words from semantically related questions. In other words, it characterize relations in related events that happen in related questions. We could find that LDA captures different relations, and the reason is that LDA describes "co-occurrence" relations because it considers words in a question. For example, people ask questions like "Is there any grocery store at Orange Beach?", and LDA is capable of capturing this kind of word relations between "grocery" and "beach" in a sentence. Thus, we think both approaches capture different semantic aspects between words.

4.3 Experiment on Labeled Question

We conducted an experiment on TST_LABEL to answer RQ2. We employed the word to word translation probabilities learnt from the parallel question-question corpus in TR, TRLM, TopicTRLM. The experimental results on metrics $P@R$, MAP, and MRR are shown in Table 5. All the results are statistically significant according to the sign test compared with the LDA.

From Table 5, we can see that LDA performs the worst. Because LDA is a coarse-grained representation to measure the relatedness between questions. It is not able to capture accurate meaning of each question. TR has better question suggestion performance compared with QL. This finding is consistent with the previous work (Jeon, Croft, and Lee 2005). The reason is that the translation model has the potential to bridge the lexical chasm between related questions. It also confirms the effectiveness of the proposed method to build parallel corpus of related questions from forum thread. TRLM has better performance than TR because TR set the probability of self-translation to 1. This introduces inconsistent probability estimates and makes the model unstable. The proposed TopicTRLM outperforms other approaches in all metrics. This confirms the effectiveness of TopicTRLM in the question suggestion task. The advantage of TopicTRLM compared with other approaches is that it fuses the latent semantic meanings of questions with lexical similarities, and this fusion promises to benefit from both the bag-of-words representation and topic model representation.

4.4 Experiment on Topics' Joint Probability Distribution

In order to answer RQ3, we conducted another experiment on TST_UNLABEL to evaluate topic level performances of the proposed method. For each queried question q, we consider its first subsequent question q' posted by the thread starter in the actual thread as its relevant result. For all the $10,000$ queried questions and their relevant results, we used the trained LDA model to infer the most probable topic. We aggregated the counts of topic transitions in the actual threads as ground truth and applied maximum likelihood estimation approach to calculate topics' joint probability using Eq. (6):

$$p(topic(q), topic(q')) = p(topic(q')|topic(q)) \times p(topic(q)). \quad (6)$$

We can get a 200×200 ($K = 200$) matrix to represent ground truth topics' joint probability distributions. In addition, for each queried question, we employed different approaches to retrieve results and considered the first result as its suggested question. We measured the difference between two probability distributions using the Kullback-Leibler divergence. The experimental results are shown in Table 6. Results in Table 6 confirm the effectiveness of the proposed TopicTRLM.

Table 6: **Comparison on difference between ground truth and methods' topics' joint probability distribution (A smaller KL-divergence value means a better performance).**

	Kullback-Leibler divergence
LDA	0.1127
QL	0.1067
TR	0.0955
TRLM	0.0911
TopicTRLM	**0.0906**

5 Conclusion and Future Work

In this paper we address the issue of question suggestion. Given a queried question, we are to suggest questions that are semantically related to the queried question and can explore different aspects of a topic tailored to users' information needs. We propose a three-step framework to tackle the problem. Specifically, we propose an effective method to build the parallel corpus of related questions from forums thread, and we propose TopicTRLM, which fuses lexical knowledge with latent semantic knowledge to measure the relatedness between questions. Extensive experiments indicate our method to build parallel corpus is effective and the TopicTRLM method outperforms other approaches.

Because we want to assist users in exploring different aspects of the topic that he/she is interested in by offering question suggestion service, it is worthwhile to investigate how to measure and how to diversify the suggested questions. Moreover, as question suggestion improves systems' understanding of users' latent intent, query suggestion for long queries might also benefit from question suggestion, which is also a future direction to investigate.

6 Acknowledgments

The work described in this paper is supported by grants from the Research Grant Council of the Hong Kong Special Administrative Region, China (Project No.: CUHK $4154/10$E and CUHK $4152/10$E). We would like to thank reviewers for their constructive suggestions. We would like to express our sincere thanks to Xinying Song and Jingtian Jiang for their help in web crawling and data extraction. We would like to thank Jie Lou for proofreading.

Words	**shore**		**park**		**condo**		**beach**	
Rank	**IBM 1**	**LDA**	**IBM 1**	**LDA**	**IBM 1**	**LDA**	**IBM 1**	**LDA**
1	shore	shore	park	park	condo	condo	beach	beach
2	beach	groceri	drive	hotel	beach	south	resort	slope
3	snorkel	thrift	car	stai	area	north	what	jet
4	island	supermarket	how	time	unit	shore	hotel	snowboard
5	kauai	store	area	area	island	pacif	water	beaver
6	condo	nappi	where	recommend	maui	windward	walk	huski
7	area	tesco	walk	beach	rent	seaport	area	steamboat
8	water	soriana	time	nation	owner	alabama	room	jetski
9	boat	drugstor	ride	tour	shore	opposit	snorkel	powder
10	ocean	mega	hotel	central	rental	manor	restaur	hotel

Table 4: **The first row shows the source words. Top** 10 **words that are most semantically related to the source word are presented according to IBM translation model** 1 **and LDA. All the words are lowercased and stemmed.**

Table 5: **Comparison on Labeled Questions (A larger metric value means a better performance).**

Metrics	**LDA**	**QL**	**TR**	**TRLM**	**TopicTRLM**
$P@R$	0.2411	0.3370	0.4135	0.4555	**0.5140**
MAP	0.3684	0.4089	0.4629	0.5029	**0.5885**
MRR	0.5103	0.5277	0.5311	0.5317	**0.5710**

References

Bernhard, D., and Gurevych, I. 2009. Combining lexical semantic resources with question & answer archives for translation-based answer finding. In *Proceedings of the 47th Annual Meeting of the ACL and 4th IJCNLP of the AFNLP*, 728–736.

Bian, J.; Liu, Y.; Agichtein, E.; and Zha, H. 2008. Finding the right facts in the crowd: Factoid question answering over social media. In *Proceeding of the 17th International Conference on World Wide Web*, 467–476.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research* 3:993–1022.

Brown, P. F.; Cocke, J.; Pietra, S. A. D.; Pietra, V. J. D.; Jelinek, F.; Lafferty, J. D.; Mercer, R. L.; and Roossin, P. S. 1990. A statistical approach to machine translation. *Computational Linguistics* 16(2):79–85.

Cao, Y.; Duan, H.; Lin, C. Y.; Yu, Y.; and Hon, H. W. 2008. Recommending questions using the mdl-based tree cut model. In *Proceedings of the 17th International Conference on World Wide Web*, 81–90.

Cao, X.; Cong, G.; Cui, B.; and Jensen, C. S. 2010. A generalized framework of exploring category information for question retrieval in community question answer archives. In *Proceedings of the 19th International Conference on World Wide Web*, 201–210.

Cong, G.; Wang, L.; Lin, C. Y.; Song, Y. I.; and Sun, Y. 2008. Finding question-answer pairs from online forums. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 467–474.

Duan, H.; Cao, Y.; Lin, C.-Y.; and Yu, Y. 2008. Searching questions by identifying question topic and question focus. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies*, 156–164.

Griffiths, T. L., and Steyvers, M. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences* 101(Suppl 1):5228–5235.

Jeon, J.; Croft, W. B.; and Lee, J. H. 2005. Finding similar questions in large question and answer archives. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, 84–90.

Och, F. J., and Ney, H. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics* 29(1):19–51.

Phan, X.-H.; Nguyen, L.-M.; and Horiguchi, S. 2008. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceeding of the 17th International Conference on World Wide Web*, 91–100.

Shrestha, L., and McKeown, K. 2004. Detection of question-answer pairs in email conversations. In *Proceedings of the 20th International Conference on Computational Linguistics*.

Wang, K.; Ming, Z.-Y.; and Chua, T.-S. 2009. A syntactic tree matching approach to finding similar questions in community-based qa services. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 187–194.

Xue, X.; Jeon, J.; and Croft, W. B. 2008. Retrieval models for question and answer archives. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 475–482.

Zhai, C., and Lafferty, J. 2004. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS)* 22(2):179–214.