Erler at the NTCIR-13 OpenLiveQ Task

Ming Chen
Wuhan, China
erlangera@whut.edu.cn

Lin Li
Wuhan, China
cathylilin@whut.edu.cn

Yueqing Sun
Wuhan, China
technology_sun@126.com

Jie Zhang
Wuhan, China
icecream@whut.edu.cn

ABSTRACT

In this paper, we present our approach to address the OpenLiveQ task at NTCIR-13 [5]. The task is a question retrieval task and can be simply defined as follows: given a query and a set of questions with their answers, return a ranked list of questions. However, there is a "gap" between queries and candidate questions, which is called lexical chasm or word mismatch problem. In our model, we improve traditional Topic inference based Translation Language Model (T^2LM) [10] by using the topic information of queries to "bridge" the gap. The translation model and the topic model are used to link different words. Experimental results show that our system reaches the competitive performance among the participants in OpenLiveQ task.

Team Name

Erler

Subtasks

OpenLiveQ

Keywords

question retrieval, translation model, topic model

1. INTRODUCTION

Community Question Answering (CQA) services have become a important alternative for online information access, such as Yahoo! Answers[1] and Quora[2]. A huge number of User Generated Content (UGC) has been accumulated in the form of Question and Answer (QA) pairs over time. Users can obtain relevant information to their search intents not only by asking questions in CQA, but also by searching for questions that are similar to their intents. Question retrieval in CQA services returns several relevant questions with possible answers directly. By this way, users do not need to wait for answers from human, which helps users save a lot of time. Therefore, the question retrieval become an important task for CQA services. For simplicity and consistency, we use the term "query" to denote a new question raised by a user and "question" to denote the answered question in the CQA archive [9].

Finding answers to questions similar to a search intent is an important information seeking strategy especially when the search intent is very specific or complicated. One of the major challenges is the lexical gap, i.e., the word mismatch between queries and candidate questions. For example, "Where can I listen to rock for free online?" and "I need a music sharing website." probably have the same meaning but in different word forms. In addition, the limited length of questions causes the sparsity of word features [3]. Therefore, traditional word frequency and document frequency statistics based retrieval models are no longer suitable for question retrieval task. Since the relationship between different words can be modeled through word-to-word translation probabilities, translation-based approaches have obtained some good results. To control the noises in translation model, some researchers introduced potential topic information in translation-based model, namely, the T^2LM [10].

In this paper, we focus on the improvement of T^2LM. We improve the T^2LM by introducing the topic information of queries. Our improved approach controls the translation noises by leveraging the topic information and balances the impact of each topic by using the topic information of query as weights. By combining both, we further improve the performance of question retrieval in CQA.

The remainder of this paper is organized as follows. Section 2 introduces the related work about question retrieval. Section 3 describes our improved retrieval model. Experiments and result analysis are reported in Section 4. Finally, conclusions are discussed in Section 5.

2. RELATED WORK

To conquer the lexical gap, researchers are constantly trying to develop more enhanced models that can bridge the chasm by linking different words. They introduced statistical machine translation model into question retrieval model. They used word-to-word translation probabilities to model the relationship between different words.

Berger et al. [1] introduced statistical translation methods to bridging the chasm in FAQ retrieval. They studied similarity calculation technique in question retrieval from the lexical level towards the semantic level. Riezler et al. [7] availed of monolingual translation based retrieval model for answer retrieval. They utilized sentence level paraphrasing approach to capture similarities between questions and answers. Xue et al. [8] presented a question retrieval model that combined a translation-based language model for the question part with a query likelihood method for the answer part.

Topic modelling based approaches, such as PLSA [4] and LDA [2], provide an elegant mathematical tool to analyze

[1]http://answers.yahoo.com
[2]http://www.quora.com

Table 1: Explanation of Terms

Term	Explanation
C	background collection
λ	smoothing parameter
$\|(q,a)\|$	word lengths of (q,a)
$tf_{w,q}$	frequency of term w in q
$P_{ml}(w\|q)$	maximum likelihood estimate of word w in q
$p(w\|t)$	probability that t is the translation of word w
$p(w\|z_i)$	distribution probability of word w under topic z_i
K	topic number

shallow semantics. Naturally, these models have attracted question retrieval researchers attention for a long time.

Zhang et al. [10] proposed a model that controls translation noise by leveraging the topic information. They focused on similarity of topic distribution between word in query and question. They utilized word distribution information under topic to improve accuracy of word-to-question similarity and further obtained better performances. But their model did not consider the topic information of query that is also valuable for question retrieval.

In this paper, we utilize the topic distribution information of queries to improve the performance of retrieval on the basis of Zhang's model [10]. We add the topic information of queries as weights into the process of word-to-question similarity to balance the impact of each topic.

3. OUR APPROACH

In this section, we give a brief introduction about the T^{2}LM [10] firstly. Then we present two our main contributions. By using the topic information of queries as weights to balance the impact of each topic, we improve the T^{2}LM. We denote the improved model as T^{2}LM* in this paper.

3.1 Topic Inference based Translation Language Model

In the T^{2}LM, given a query $query$ and a QA pair (q,a) consisted of a question q and an answer a, a ranking score $P(query|(q,a))$ is computed as follows:

$$P(query|(q,a)) = \prod_{w\in query}\left(\frac{|(q,a)|}{|(q,a)|+\lambda}P_{t^2lm}(w|(q,a)) + \frac{\lambda}{|(q,a)|+\lambda}P_{ml}(w|C)\right) \tag{1}$$

$$P_{t^2lm}(w|(q,a)) = \mu_1 P_{ml}(w|q) + \mu_2\sum_{t\in q}(p(w|t)P_{ml}(t|q)) + \mu_3\sum_{t\in q}\left(P_{ml}(t|q)\sum_{i=1}^{K}(p(w|z_i)p(t|z_i))\right) + \mu_4 P_{ml}(w|a) \tag{2}$$

$$P_{ml}(w|q) = \frac{tf_{w,q}}{|q|} \qquad P_{ml}(w|a) = \frac{tf_{w,a}}{|a|} \qquad P_{ml}(w|C) = \frac{tf_{w,C}}{|C|} \tag{3}$$

The explanations of terms in Equation 1 to 3 are showed in Table 1. w is a word in query $query$, and t is a word in question q. $|q|$, $|a|$, $|C|$ have similar meanings to $|(q,a)|$; $tf_{w,a}$ and $tf_{w,C}$ have similar meanings to $tf_{w,q}$; $P_{ml}(w|a)$ and $P_{ml}(w|C)$ have similar meanings to $P_{ml}(w|q)$; $p(t|z_i)$ has a similar meaning to $p(w|z_i)$; and μ_1, μ_2, μ_3 and μ_4 balance the impact of each component and $\mu_1+\mu_2+\mu_3+\mu_4=1$.

3.2 Topic Information of Query

From the Equation 2 we can see that the significance of each topic is equal in T^{2}LM. And the word topic distribution probabilities from static corpus are used statically for a dynamic query. Although the diverse topic information of various queries is beneficial for question retrieval generally, the information is ignored. In this paper, we propose an approach to exploit this topic information. In our approach, the topic information of queries is used as weights of each topic to improve the process of capture word-to-question similarity. More formally, given a query $query$ and a topic z_i, $P(query|z_i)$ denoting the weight of topic z_i for query $query$ is computed as follows:

$$P(query|z_i) = \frac{\prod_{w\in query}p(w|z_i)}{\sum_{j=1}^{K}\prod_{w\in query}p(w|z_j)} \tag{4}$$

Here w is a word in query $query$; $p(w|z_i)$ and $p(w|z_j)$ are the distribution probability of word w under topic z_i and z_j; and K is topic number. The denominator in Equation 4 may be zero in some cases. To solve the problem, we make a compromise that we set $P(query|z_i)=1/K$ for each topic z_i if the problem happens. Then the $P(q|z_i)$ is used as weights of each topic in the process of capturing word-to-question similarity. The specific method is showed as follows:

$$P_{t^2lm^*}(w|(q,a)) = \mu_1 P_{ml}(w|q) + \mu_2\left(\sum_{t\in q}P(w|t)P_{ml}(t|q)\right) + \mu_3\left(\sum_{t\in q}P_{ml}(t|q)K\sum_{i=1}^{K}(P(query|z_i)p(w|z_i)p(t|z_i))\right) + \mu_4 P_{ml}(w|a) \tag{5}$$

Here w is a word in query $query$. By multiplying K we control the range of topic part unchanged so that scope of μ_3 is the same as before. We denote the improved model as T^{2}LM* in this paper.

4. EXPERIMENTS AND EVALUATION

4.1 Setup

The OpenLiveQ task at NTCIR-13 provides a data set from Yahoo! Chiebukuro search query log. It contains 2000 queries, among them 1,000 queries for training and the rest for testing. For each query, the data set provides 1000 questions to be ranked. For each question, the data set provides information about question includes query ID, question ID, title, body of the question and body of the best answer for the question etc. In addition, the data set provides click-through data for some of the questions. The size of the data set is 2,000 queries, 1,967,274 questions and 440,163 click-through data. In this paper, we used the queries, questions and their answer.

For data preprocessing, we used Mecab to segment words and remove the stop words for each question pair in the data set firstly. We use the GIZA++ toolkit [6] for learning the IBM Translation Model 1 to get the word-to-word

Table 2: Offline test result

Model	nDCG@10
T2LM*	0.39139
T2LM	0.3867
TLM	0.37304
TM	0.37985
Demo	0.40566

translation probabilities. We pool the QA pairs and the answer-question pairs together as the input to this toolkit [8]. We get a word-to-word translation probability list after training. For topic model part, we think of questions as documents and utilize LDA [2] model to model question set. We set the topic number as 70.

We use three retrieval models, the Topic Model (TM), the Translation-based Language Model (TLM) and the Topic Inference-based Translation Language Model (T^2LM), as baseline methods. We conduct experiments to demonstrate the effect of our proposed model in Section 3, T^2LM*.

4.2 Results

In NTCIR-13 OpenLiveQ task, we have submitted 12 runs. Table 2 shows the best result for each model in offline test in terms of nDCG@10. From the table, we can see that among the three baseline models, T^2LM performs the best and TM performs the worst; TLM between TM and T^2LM. And our approach T^2LM* obtains the best performance in four models. However, we can know that the demo model, learning to rank model, performs better than our model. And in the offline tests we have made fifth place in all participating teams.

The reasons for this result may be many, but may be mainly the following. The T^2LM performs better than the TLM and TM, because the T^2LM combines with the advantages of TLM and TM. The results are consistent with the reported in previous work [10]. The T^2LM* performs better than the T^2LM, and the UT^2LM performs better than the T^2LM*. The underlying reasons are that the T^2LM* utilizes the topic information of query as weight to improve the topic component on the basis of T^2LM. But we can see the improvements are very small. We think that's because the query is too short. The topic information of queries are very sparse. As for the results of Demo better than our model, we think it is because the demo model uses the data set to provide all the information, and our model only use the body of question and the body of its answer. Obviously the other information including last update time of the question, number of answers for the question and category of the question is helpful to optimize the retrieval results. On the other hand, training of the translation model in our model can be further optimized. In this task, we use the QA pairs and the answer-question pairs as parallel corpus to train the translation model. If better corpus is used as training data, retrieval results of our model can be further improved.

Figure Figure 1 shows the results of online test. We can see that compared to other teams we have achieved good results. In addition to ORG and YJRSI, our approach has been better than other teams in most of the tests.

5. CONCLUSIONS

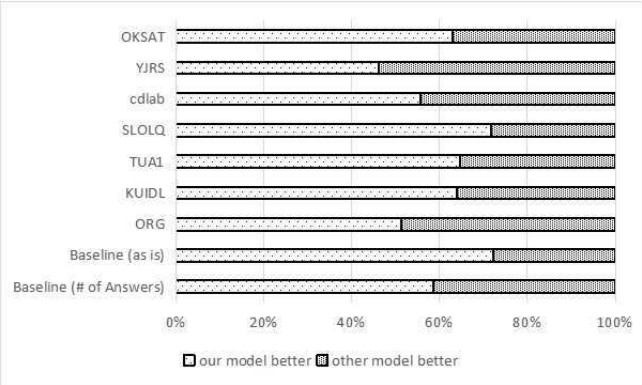

Figure 1: Online test results

Question retrieval is an important component in Community Question Answering (CQA) services. In this paper, we propose a novel approach by first using topic information of query to improve the T^2LM. We apply it to the OpenLiveQ task at NTCIR-13, and achieve satisfactory results.

For the future work, we will continue out better features and models to improve question retrieval result.

6. REFERENCES

[1] A. Berger, R. Caruana, D. Cohn, D. Freitag, and V. Mittal. Bridging the lexical chasm: statistical approaches to answer-finding. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 192–199. ACM, 2000.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[3] L. Chen, J. M. Jose, H. Yu, F. Yuan, and D. Zhang. A semantic graph based topic model for question retrieval in community question answering. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 287–296. ACM, 2016.

[4] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine learning*, 42(1):177–196, 2001.

[5] M. P. Kato, T. Yamamoto, and T. Manabe. Overview of the ntcir-13 openliveq task. In *Proceedings of the NTCIR-13 Conference*, 2017.

[6] F. J. Och and H. Ney. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 440–447. Association for Computational Linguistics, 2000.

[7] S. Riezler, A. Vasserman, I. Tsochantaridis, V. Mittal, and Y. Liu. Statistical machine translation for query expansion in answer retrieval. In *Annual Meeting-Association For Computational Linguistics*, volume 45, page 464, 2007.

[8] X. Xue, J. Jeon, and W. B. Croft. Retrieval models for question and answer archives. In *Proceedings of the 31st annual international ACM SIGIR conference on*

Research and development in information retrieval, pages 475–482. ACM, 2008.

[9] W. N. Zhang, Z. Y. Ming, Y. Zhang, T. Liu, and T. S. Chua. Capturing the semantics of key phrases using multiple languages for question retrieval. volume 28, pages 888–900, 2016.

[10] W. N. Zhang, Z. Yu, and T. Liu. A topic inference based translation model for question retrieval in community-based question answering services. volume 38, pages 313–321, 2015.