Erler at the NTCIR-13 OpenLiveQ Task

Ming Chen, Lin Li, Yueqing Sun, Jie Zhang

School of Computer Science and Technology, Wuhan University of Technology, China

Introduction

Community Question Answering (CQA) services have become a important alternative for online information access, such as Yahoo! Answers and Quora. Question retrieval is an important task for CQA services. The task can be simply defined as follows: given a query and a set of questions with their answers, return a ranked list of questions.

A major challenge of the task 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.

Fig. 1. Question Retrieval

We focus on using Translation Model and Topic Model to improve retrieval model. In our model, Translation Model and Topic Model "bridge" the word gap by linking different words. Besides, we use topic information of query improve retrieval result further.

Our Approach

Our model can be outlined in **Fig. 2**. We train Topic Model and Translation Model to get word topic distribution probability $P_{to}(w|z_i)$ and translation probability $P_{tr}(w|t)$. They are used to extract topic information of query and calculate likelihood. **Fig. 3** and **Fig. 4** show two examples of $P_{to}(w|z_i)$ and $P_{tr}(w|t)$. In this task, we use question and answer pairs as training data.

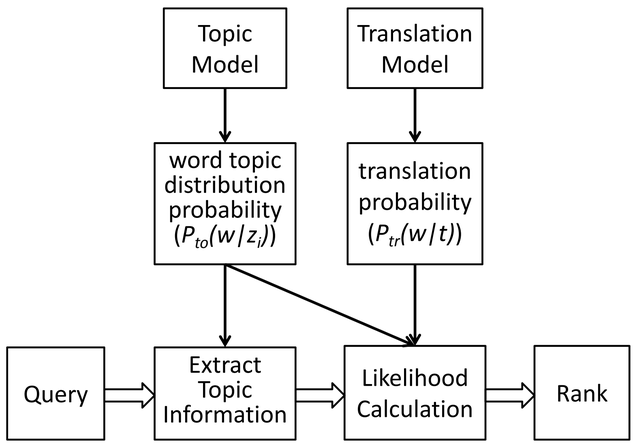

Fig. 2. Our Approach

Topic 1		Topic 2	
w	$P_{to}(w\|z_1)$	*w*	$P_{to}(w\|z_2)$
する	0.014324	大学	0.047982
家	0.012173	就職	0.013081
あり	0.011606	高校	0.012793
い	0.011130	合格	0.010597
いる	0.010675	受験	0.009892
工事	0.009775	偏差値	0.009028
部屋	0.009184	学科	0.008651
業者	0.008671	学生	0.008585
電気	0.007376	者	0.008320
設置	0.007285	進学	0.007968

Fig. 3. An example of $P_{to}(w|z_i)$

t	$P_{tr}(w\|t)$	*t*	$P_{tr}(w\|t)$
あり	0.026	設備	0.008
よう	0.015	電気工事	0.008
電力	0.013	冷蔵庫	0.007
機械	0.011	電	0.007
用	0.011	配線	0.007
物	0.011	工事	0.006
エアコン	0.010	これから	0.006
工学部	0.009	節約	0.006
家	0.009	電子	0.006
暖房	0.009	ブレーカー	0.005

Fig. 3. An example of $P_{tr}(w|t)$ where *w*="電気"

In our model, we use the topic information of *query* as weights to balance the impact of each topic. *K* is topic number. $P(query|z_i)$ comes from thematic information extracted from *query*:

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

Then we can get $P_{to}(w|t,query)$ where $w \in query$:

$$P_{to}(w|t,query)=\sum_{i=1}^{K}\left(P(query|z_i)P_{to}(w|z_i)P_{to}(t|z_i)\right)$$

Query likelihood is a generative model that assumes that the question answer pair *(q, a)* is a sample of a multinomial distribution of terms [2, 3]. *(q, a)* are ranked according to the probability they generate the *query*. We estimate this probability by interpolating the term distribution in the *(q, a)* with the term distribution in the collection:

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

Here we use length of *(q, a)* as smoothing parameter. $P_{ml}(w|C)$ is the distribution of word *w* in the collection *C*. And $P(w|(q, a), query)$ derived based on Language Model, Translation Model and Topic Model:

$$P(w|(q,a),query)=\mu_1 P_{ml}(w|q)+\mu_2\sum_{t\in q}\left(P_{tr}(w|t)P_{ml}(t|q)\right)+\mu_3\sum_{t\in q}\left(P_{to}(w|t,query)P_{ml}(t|q)\right)+\mu_4 P_{ml}(w|a)$$

Here we use μ_1, μ_2, μ_3 and μ_4 balance the impact of each component and $\mu_1+\mu_2+\mu_3+\mu_4=1$.

Experiments and Conclusions

We use three retrieval models, the Topic Model (TM) [3], the Translation-based Language Model (TLM) [4] and the Topic Inference-based Translation Language Model (T²LM) [5], as baseline methods. We conduct experiments to demonstrate the effect of our proposed model T²LM*. The **Fig. 5** shows the best result for each model in offline test in terms of nDCG@10.

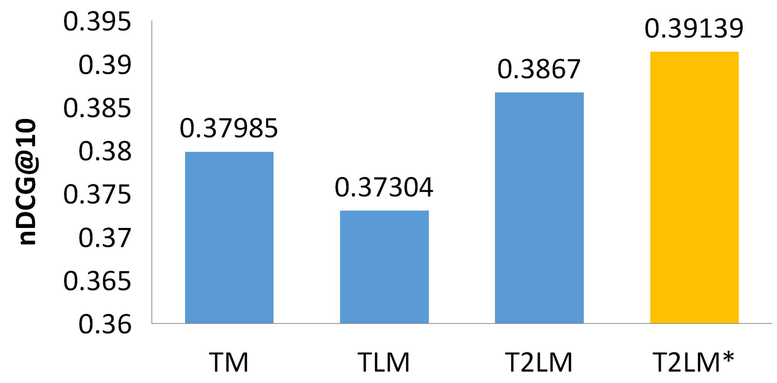

Fig. 5. Offline Test Results

T²LM performs better than TLM and TM, because it combines the advantages of TLM and TM. And our model T²LM* performs better than T²LM. The underlying reasons are that the T²LM* utilizes the topic information of query as weight to improve the topic component on the basis of T²LM. But we can see the improvements are very small. We think that's because the query is too short. In this task, each query contains 1 or 2 words. The topic information of queries are too sparse to further improve the search results.

The improvement achieved by our method is not obvious. And in the offline test we have made fifth place in all participating teams. The reason may be that our model only use the content of the question and its answer. Obviously the other information including last update time of the question and category of the question in the data set is helpful to optimize the retrieval results.

References

[1] Ponte, J.M., Croft, W.B.: A language modeling approach to information retrieval. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 275–281 (1998)

[2] Jeon, J., Croft, W.B., Lee, J.H.: Finding similar questions in large question and answer archives. In: CIKM, pp. 84–90 (2005)

[3] Wei, W., Croft, W.B.: LDA-based document models for ad-hoc retrieval. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 178–185 (2006)

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

[5] Zhang, W.N., Zhang, Y., Liu, T.: A topic inference based translation model for question retrieval in community-based question answering services. Chin. J. Comput. 38(2), 313–321 (2015)

Acknowledgements

The authors wish to acknowledge the support from the NTCIR Project and the organizers of OpenLiveQ task.

Contacts

1. Ming Chen, erlangera@whut.edu.cn
2. Lin Li, cathylilin@whut.edu.cn