

# 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

Wuhan, China

cathylilin@whut.edu.cn



01

Introduction

02

Our Model

03

Experiments

04

Conclusions



# Introduction

What problem do we have to solve?

1

# Introduction

Search

How to update ubuntu to the latest version?

684 results

**relevance** newest votes active

2 votes  
3 answers  
Q: How to update VLC to the latest version on Ubuntu  
Related to: How to update VLC to the latest version? Except that there are no builds for Natty in this VLC ppa: ppa:videolan/stable-daily Index of /videolan/stable-daily/**ubuntu**/dists: oneiric precise So **how** do I **update** VLC (currently 2.0) on **Ubuntu** 11.04, Natty Narwhal? ...  
ppa vlc asked Feb 19 '12 by Kuz Mitch

1 vote  
1 answer  
Q: upgrade ubuntu to the latest released version  
: 14.04 Codename: trusty But i think this is not **the latest version** of **the Ubuntu** released on **the Ubuntu** site.when i exec sudo **update**-manager -di get nothing **to update!** **How** can i upgrade my system **to the latest version** without losing any application or data ... I want **to** upgrade my **Ubuntu to the latest** version.My **ubuntu** release when i run **the** lsb\_release -a is: No LSB modules are available. Distributor ID: **Ubuntu** Description: **Ubuntu** 14.04.3 LTS Release ....  
upgrade updates asked Oct 24 '15 by Emad Helmi

2 votes  
1 answer  
Q: How to update Apache2 on Ubuntu 14.04 Server to the latest version?  
and it's marked as High Risk. I've been searching for over an hour on **how to update** my Apache but **to** no avail. I've searched for **the latest version**, it's 2.4.10 but I have no idea **to** "install it" or ... **update** it, or patch it. I've done **the** apt-get **update**/upgrade 10 times, Apache stays **the** same. **The** OS is **Ubuntu** 14.04 Server 64bit. Please help! ...  
server updates apache2 asked Oct 20 '14 by Alex Iordache

Figure 1: An example of Question Retrieval

## Task:

The task was simply defined as: given a query and a set of questions with their answers, return a ranked list of questions.

## Challenge:

People expresses similar meanings through different words.

## Example:

update/upgrade  
更改/更新

We need to model the relationship between different terms to improve the retrieval model.

# Introduction

## Task

Question retrieval is an important task for Community Question Answering services.



## Challenge

The lexical gap, the word mismatch between queries and candidate questions.

## Solution

We propose a retrieval model based on Translation Model and Topic Model.

## Example

“I need a music sharing website.”  
“Where can I listen to rock for free online?”

# Introduction

Word	Translation probability	Word	Translation probability
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

Table 1: An example of Translation Model (source word is “電気”)

## Solution:

- We utilize Translation Model to model the relationship between different words;
- we use **translation probability** concretely.

# Introduction

Topic 1		Topic 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

Table 2: An example of Topic Model

## Solution:

- Similarly, we can use **word topic distribution probability** from Topic Model.
- The two words get a **higher correlation** if they have higher probabilities of distributions under a certain topic.

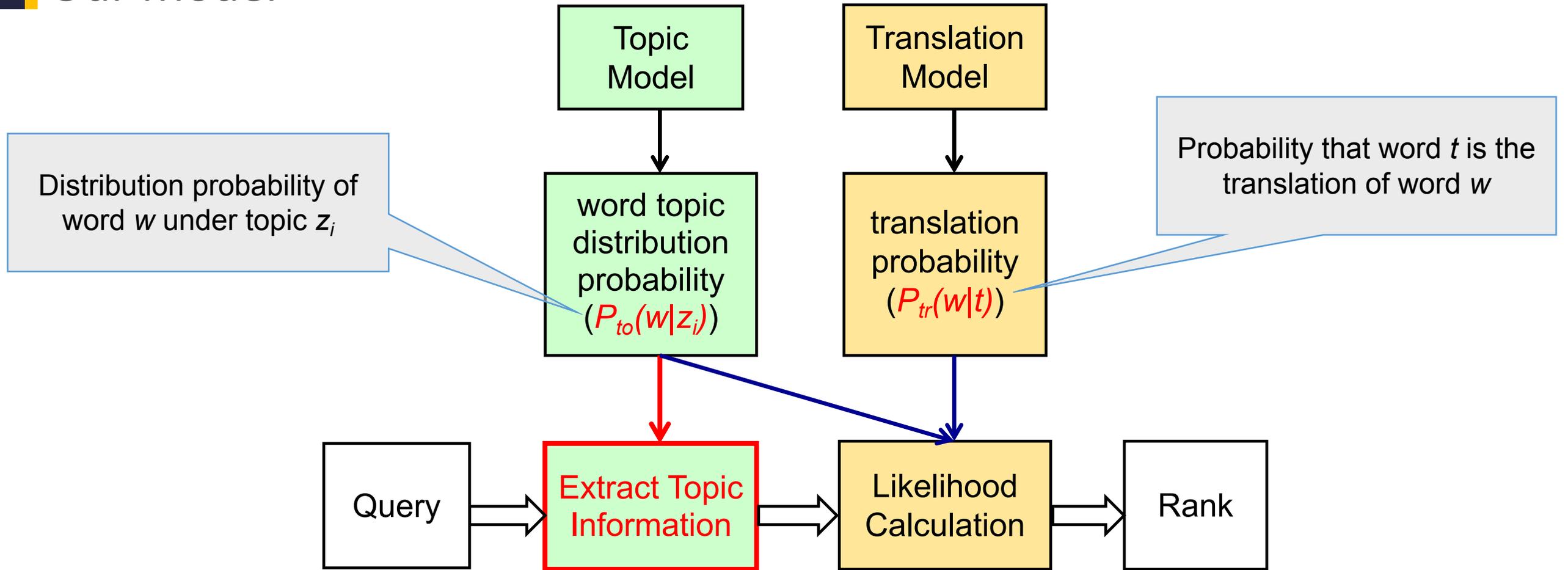


## Our Model

How to combine Translation Model and Topic Model to improve retrieval model?

2

# Our Model



Ming Chen, Lin Li, Qing Xie, Translation Language Model Enhancement for Community Question Retrieval Using User Adoption Answer, *APWEB-WAIM 2017*: 251-265.

# Our Model

## ✓ Translation Model

- Statistical Machine Translation
- The Noisy Channel Model
- Expectation Maximization (EM) Algorithm
- Translation probability  $P_{tr}(w | t)$
- Monolingual Parallel Corpus

## ✓ Topic Model

- Latent Dirichlet Allocation (LDA) Model
- Word topic distribution probability  $P_{to}(w | z_i)$

# Our Model

## Likelihood Calculation

Query likelihood is a generative model that assumes that the question answer pair  $(q, a)$  is a sample of a **multinomial distribution** of terms. We estimate this probability by interpolating the term distribution in the  $(q, a)$  with the term distribution in the collection:

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

Here  $P_m(w | C)$  is the distribution of word  $w$  in  $C$ .

$C$  is the training collection.

We use length of  $(q, a)$  as a smoothing parameter.

# Our Model

## Likelihood Calculation

Translation Model

$$P((w, query) | (q, a)) = \mu_1 P_m(w | q) + \mu_2 \sum_{t \in q} (P_{tr}(w | t) P_m(t | q))$$
$$+ \mu_3 \sum_{t \in q} \left( \sum_{i=1}^K (P(query | z_i) P_{to}(w | z_i) P_{to}(t | z_i)) P_m(t | q) \right) + \mu_4 P_m(w | a)$$

Topic Model

Here  $P_m(w | q)$  is the distribution of word  $w$  in  $q$ .

We use  $\mu_1, \mu_2, \mu_3$  and  $\mu_4$  balance the impact of each component and  $\mu_1 + \mu_2 + \mu_3 + \mu_4 = 1$ .

Example:

$w = \text{"A"}; q = (\text{"A"}, \text{"B"}, \text{"C"}); P_m(w | q) = 1/3$

# Our Model

## Extract Topic Information of a Query

For different queries we can get different weights of each topic as follows:

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

$K$  is the number of topics.

To balance the impact of each topic, which is different from traditional model.

### Example:

query = (“大学”, “偏差值”)

$$P(query|Topic1) = \frac{0.047982 \times 0.009028}{0.047982 \times 0.01918 + 0.009028 \times 0.006282} \approx 0.7824$$

Word	Topic 1	Topic 2
大学	0.047982	0.01918
偏差值	0.009028	0.006282



# Experiments

How do we conduct experiments?

3

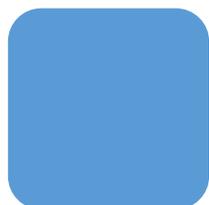
# Experiments

## Baselines



### **TM (Topic-based Model)**

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)



### **TLM (Translation-based Language Model)**

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)



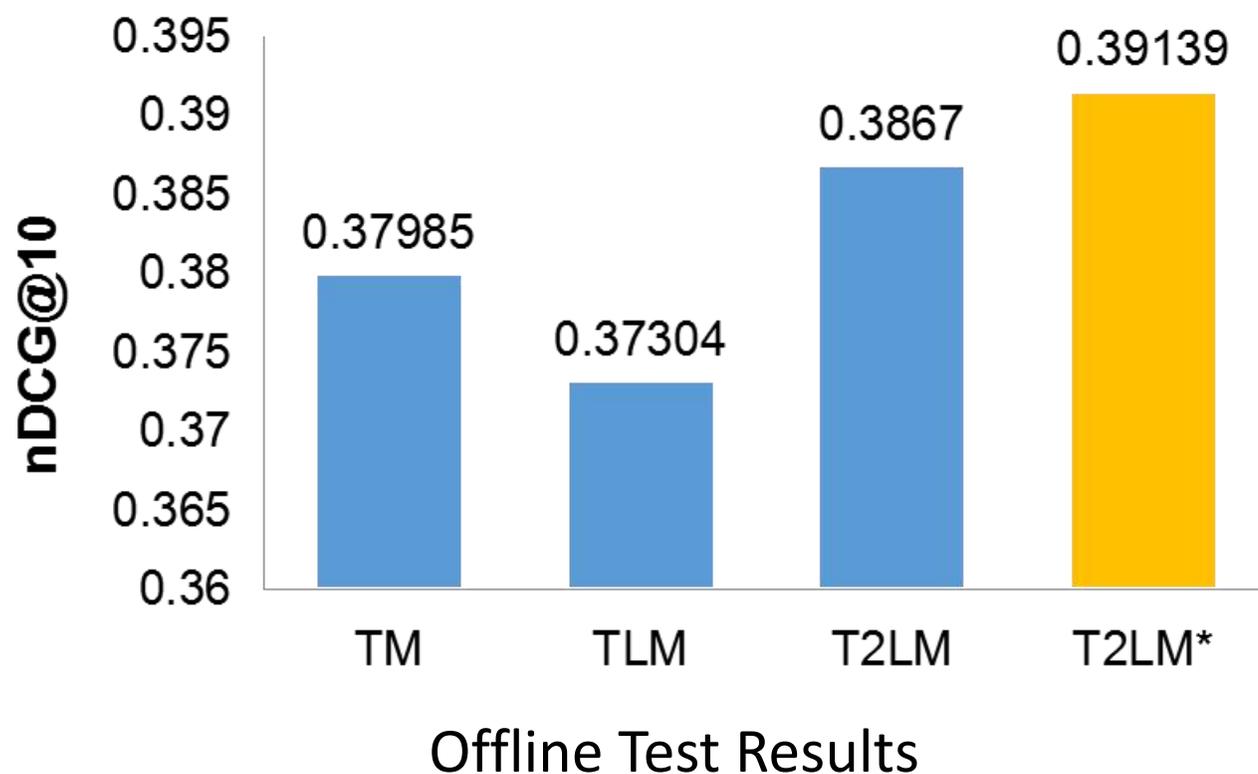
### **T<sup>2</sup>LM (Topic Inference-based Translation Language Model)**

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)

# Experiments

## Experimental results

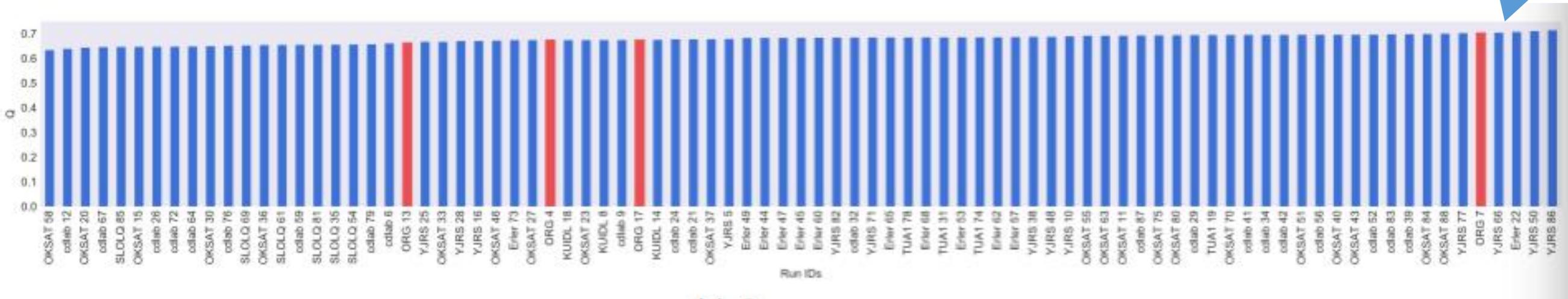
In terms of nDCG@10, **T<sup>2</sup>LM\*** performs best among traditional topic and translation models.



# Experiments

## Experimental results

Better than baseline in terms of Q measure.



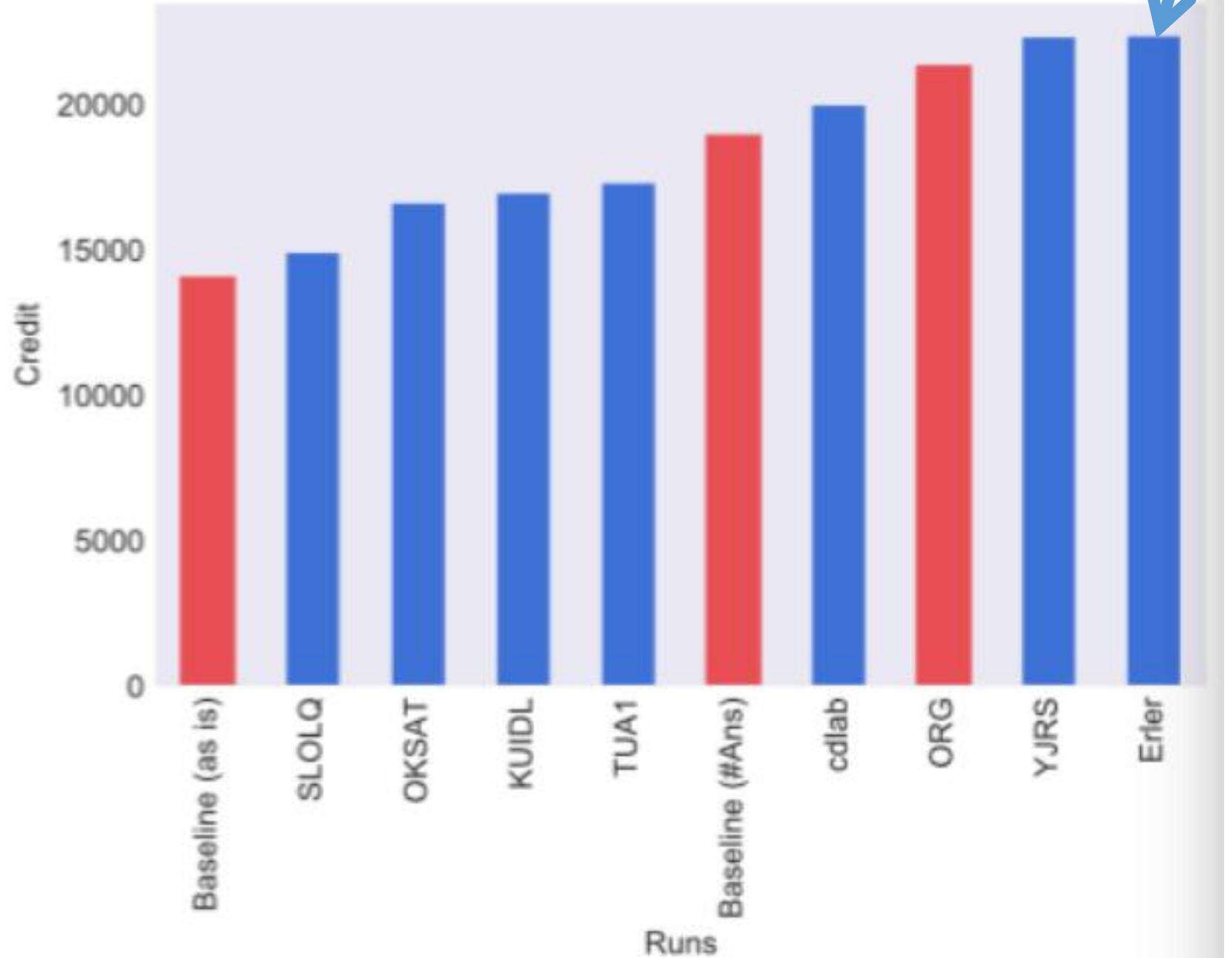
Offline Test Results



# Experiments

## Experimental results

Cumulated credits in the online evaluation.





# Conclusions

What can we get?

4

# Conclusions

1

We propose a novel approach by using the topic information of a query to improve the likelihood calculation.

---

2

Experiments on OpenLiveQ task demonstrate the effectiveness of the proposed retrieval model.

---

3

In the online test, our team and YJRS team have been tied for the first place.

1

## Integrating other information into our model

- Our model only use the content of the question and its answer.
- Obviously the other information including last update time of the question and the category of a question is helpful to optimize the retrieval results.

2

## Looking for better training corpus

- In this task, we use the QA pairs and the answer-question pairs as parallel corpus to train the translation model.
- Training of the translation model in our model can be further optimized.



*Thank you for  
listening!*

