# THUIR at the NTCIR-16 WWW-4 Task

Shenghao Yang, Haitao Li, Zhumin Chu, Jingtao Zhan, Yiqun Liu*, Min Zhang, and Shaoping Ma

Department of Computer Science and Technology, Institute for Artificial Intelligence,
Beijing National Research Center for Information Science and Technology,
Tsinghua University, Beijing 100084, China

yiqunliu@tsinghua.edu.cn

# Introduction

- WWW-4, an ad hoc search
- In the WWW-4 task. We adopt several re-ranking methods based on the retrieval result of BM25:
  - (1) list-wise learning-to-rank methods： LambdaMART and Coordinate Ascent.
  - (2) PROP： a popular pretrained language model tailored for information retrieval.
  - (3) BERT-Prompt： a BERT model tuned with prompt learning to align pre-training and fine-tuning for better performance.

# Methods

## Learning-to-Rank methods

### Dataset

- Train data
  - MQ2007 and MQ2008
  - About 1700 + 800 topics
- Validation data
  - WWW1-3 English test set
  - 260 topics

### Features Extraction

- **Preprocessing:** lowercasing, tokenization, removing stop words, and stemming
- **Four feature fields:** whole document, anchor text, title, and URL
- **Eight features:**term frequency (TF), inverse document frequency (IDF), TF * IDF, document length (DL), BM25, LMIR.ABS, LMIR.DIR and LMIR.JM.

**Ranklib:**

LambdaMART(run4) & Coordinate Ascent (run5)

# Methods

## PROP

### Dataset

- WWW1-3 English test set
- For all 260 topics, we divide the training set and validation set in the ratio of 4:1.
- we convert the labeled documents of each topic into <topic,doc1,doc2> tuples.

### Training

- we concatenate the query and document as input of PROP.
- Select the model checkpoints:
  - First method: select the best performing checkpoint on a validation set. (run1)
  - Second methods: inherit the tuned hyperparameters of run1 but use the full labeled data for training without validation. (run2)

# Methods

## BERT-Prompt

### Training

- We feed BERT in the "[query] [mask] [document]" format and predict the probability of all words in the vocabulary at the [mask].
- We utilize "yes", "and", and "so" as the positive words. "but", "yet", and "however" are the negative words.
- The relevance score equals the subtraction results between the average probability values of positive words and negative words.

# Experimental Results

**Table 1: WWW-4 official results of THUIR runs based on the gold relevance assessments.**

| Run | Model | Mean nDCG | | Mean Q | | Mean nERR | | Mean iRBU | |
|---|---|---|---|---|---|---|---|---|---|
| THUIR-E-CO-NEW-1 | PROP | 0.3596 | 5 | 0.2931 | 2 | 0.5102 | 4 | 0.7449 | 7 |
| THUIR-E-CO-NEW-2 | PROP | 0.3670 | 3 | 0.2944 | 1 | 0.5289 | 1 | 0.7544 | 5 |
| THUIR-E-CO-NEW-3 | BERT-Prompt | 0.3222 | 13 | 0.2494 | 14 | 0.4281 | 18 | 0.7166 | 16 |
| THUIR-E-CO-NEW-4 | LambdaMart | 0.3094 | 16 | 0.2288 | 16 | 0.4672 | 13 | 0.7510 | 6 |
| THUIR-E-CO-NEW-5 | Coordinate Ascent | 0.3405 | 6 | 0.2667 | 8 | 0.4783 | 9 | 0.7545 | 4 |

- PROP achieves the best overall performance on four evaluation metrics.
- BERT-Prompt underperforms PROP and performs comparably with the learning-to-rank methods.
- We speculate that the used prompt learning approach is relatively simple.

# Case Study

- The neural ranking models can capture the semantic similarity between topic and document, while the learning-to-rank methods focus on lexical features and may encounter the "semantic gaps" problem.

**Table 2: Case study on topic 217 and document 2bd06c76-f5f3-4be1-98c8-0d66dbdf41a6. In this case, the high idf topic term "inventor" appears less frequently. It results in a poor ranking performance of learning-to-rank method. While neural ranking model can capture semantic relevance rather than lexical matching and performs better. Specifically, the relevant document in this case is ranked by PROP at the first position, while Coordinate Ascent ranks it at the 77th position.**

| |
|---|
| Topic: inventor of the Web<br>Description: Who is the inventor of the World Wide Web?<br>Rank result: PROP(1) Coordinate Ascent(77) Gold relevance judgment(L2) |
| https://www.famous**inventor**s.org/tim-berners-lee tim berners-lee \| biography, inventions and facts famous **inventor**shome about blog contacttim berners-lee tim berners-lee invented "world wide **web**" and "html". tim berners-lee (formally sir timothy john berners-lee) is the **inventor** of the world wide **web**. he was born in britain on june 8th, 1955 and graduated from oxford university with a first class honors degree in physics. he was influenced by his parents' interest in computers and technology, as they were part of the team who built the first commercial computer. as a child he was deeply interested in trains and took them apart to learn how they worked. at college he built his first computer using an old television, a soldering iron and a processor. |

# Case Study

- Although the neural ranking model has better semantic matching ability, it is limited by the input length (512) of transformer model

**Table 3: Case study on topic 225 and document 2250aaa2-6a41-40a4-8924-72c782129ec9. In this case, the topic terms appears at later positions in the document and the topic terms "signifier" and "saussure" even not appears in following table. Limited to the input length, PROP can't obtain enough relevant information and performs poorly. While, learning-to-rank method is not limited by the input length and performs better. Specifically, the relevant document in this case is ranked by PROP at the 53th position, while Coordinate Ascent ranks it at the 4th position.**

| Topic: signifier saussure theory<br>Description: You want to know the meaning of term "signifier" in linguist Saussure's theory<br>Rank result: PROP(53) Coordinate Ascent(4) Gold relevance judgment(L2) |
|---|
| https://www2.slideshare.net/mattheworegan/stuart-hall-representation-**theory** stuart hall - representation **theory** slideshare uses cookies to improve functionality and performance, and to provide you with relevant advertising. if you continue browsing the site, you agree to the use of cookies on this website. see our user agreement and privacy policy. slideshare uses cookies to improve functionality and performance, and to provide you with relevant advertising. if you continue browsing the site, you agree to the use of cookies on this website. see our privacy policy and user agreement for details.slideshareexploresearchyouupload login signupsubmit search home explore successfully reported this slideshow. we use your linkedin profile and activity data to personalize ads and to show you more relevant ads. |

# Case Study

- Complete topic expression and distinct terms may help the model to determine relevant information more accurately

**Table 4: Case study on topics where both PROP and Coordinate Ascent performs well or poorly.**

| topic ids | topic | Mean nDCG | | Mean Q | |
|---|---|---|---|---|---|
| | | PROP | CA | PROP | CA |
| 201 | Timnit Gebru Google | 0.8002 | 0.8002 | 0.8435 | 0.7488 |
| 245 | chicken breast recipes | 0.8880 | 0.9450 | 0.9075 | 0.9524 |
| 250 | trek emonda price | 0.9653 | 0.9351 | 0.9888 | 0.9896 |
| 203 | idf inventor | 0 | 0 | 0 | 0 |
| 206 | DC and Marvel characters | 0 | 0 | 0 | 0 |
| 207 | dirty loops bassist | 0 | 0 | 0 | 0 |
| 220 | half life | 0 | 0.0367 | 0 | 0.0095 |

# Conclusion and Future Work

- We investigate PROP, BERT-based neural ranking models, and learning-to-rank methods. Experimental results show the importance of pre-training.
- In the future, we will try to further optimize our prompt learning design approach to further improve the ranking performance of BERT-Prompt.


TSINGHUA UNIVERSITY
1911


Information Retriever @ Tsinghua University

# Thanks!

# Q&A