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

❖ Introduction

- **We participated in English task in WWW-4 task.We adopt several re-ranking models：**
  - list-wise learning-to-rank methods
  - PROP： a popular pre- trained language model tailored for information retrieval
  - BERT-Prompt: a BERT model tuned with prompt learning to align pre-training with fine-tuning for better performance.

❖ English Task

- **Learning to rank(New runs)**
  - We adopt MQ2007 and MQ2008 as the training data.
  - We extracted features in four different fields: whole document, anchor text, title, and URL. Each field contains eight different features.
  - We extract the following eight features in four fields: term frequency (TF), inverse document frequency (IDF), TF * IDF, document length (DL), BM25, LMIR.ABS, LMIR.DIR and LMIR.JM.
  - We choose Lambdamart and Coordinate Ascent as the final submission because these models perform well in the validation set.
- **PROP(New runs)**
  - we concatenate the query and document as input of PROP, The output embedding of [CLS] token is fed into a linear layer to obtain the relevance scores of the query-document pairs
  - We train PROP on the dataset collected from WWW 1-3. For all 260 topics, we divide the training set and validation set in the ratio of 4:1.
- **BERT-Prompt(New runs)**
  - we feed BERT in "[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

- PROP achieves the best overall performance on four evaluation metrics.
- BERT-Prompt underperforms PROP and performs comparably with the learning-to-rank methods.
- Prompt learning does not substantially improve the ranking performance as expected. We speculate that the used prompt learning approach is relatively simple.

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

❖Case Study

- we choose PROP and Coordinate Ascent as neural ranking model and learning-to-rank method for the comparison.
- The neural ranking models can capture the semantic similarity between topic and document, but it is limited by the input length (512) of transformer model. The learning-to-rank methods do not suffer from this limitation but may encounter the "semantic gaps" problem.
- Complete topic expression and distinct terms may help the model to determine relevant information more accurately

❖Conclusion & 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.