RUCIR at the NTCIR-15 WWW-3 Task

Xiaochen Zuo
Renmin University of China
zuoxc@ruc.edu.cn

Jing Yao
Renmin University of China
jing_yao@ruc.edu.cn

Zhicheng Dou
Renmin University of China
dou@ruc.edu.cn

ABSTRACT

The RUCIR team participated in both Chinese and English subtasks of the NTCIR-15 We Want Web-3 (WWW-3) task. This paper describes our approaches and results in both subtasks. In the Chinese subtask, we use Bert [2] on the SogouQCL [17] dataset and a commercial dataset. In English subtask, we use Bert and learning to rank method on TREC Web Track dataset and MS MARCO Passage Ranking dataset [1]. Our approaches achieved the best performances in Chinese subtask.

KEYWORDS

ad-hoc search, document ranking, learning to rank

TEAM NAME

RUCIR

SUBTASKS

Chinese and English

1 INTRODUCTION

The goal of ad-hoc retrieval is to rank candidate documents according to the query given by the user, and there are plenty of methods designed for this task. Most of traditional methods are based on exact matching signal between words from query and document, such as BM25 [10]. With the widespread application of machine learning algorithms, some feature-based methods have been proposed to address ad-hoc search problem, including RankNet [1], AdaRank [16] and LambdaMART [14], etc. Compared to these traditional machine learning methods, deep neural ranking models can better extract the semantic features of text. Neural ranking models can be classified into two categories: representation-focused model and interaction-focused model. Representation-focused models learn the representations of query and document separately and then calculate the similarities between them, including ARC-I [3], DSSM [4] and etc. Interaction-focused models employ interaction between query and document first, and then design neural networks to learn matching patterns based on the result of interaction. This type of model includes ARC-II [3], KNRM [15] and MatchPyramid [9], etc. Additionally, Some models combine the advantages of these two types of models, such as Duet [8].

Recently, Bert [2] has achieved remarkable results in many natural language processing tasks. It can better process sequence information and effectively extract contextual semantic information in sentences. Bert is also effective for ad-hoc retrieval task. When using Bert, we adopt an idea similar to interaction-focused models. Specifically, we concatenate the query and the document through a separator, then train the sequence through Bert. Finally, we use the output vector at the first position to calculate a ranking score. In the Chinese subtask, in order to improve the learning effect of the model, we use the idea of multi-task learning to calculate two scores at the same time. One score is used to predict manual annotations and the other is used to predict PSCM [13] score.

The rest of the paper is organized as follows. Section 2 introduces our approaches for the task. Section 3 introduces the datasets for two subtasks and shows evaluation results of our submitted runs and provides discussion. Finally we conclude in Section 4.

2 OUR APPROACHES

In this section, we introduce our approaches on WWW-3 task. We use Bert on both Chinese subtask and English subtask. However, due to the differences in the characteristics of Chinese dataset and English dataset, we utilize Bert in different ways for the two subtasks, and we will describe the differences in the following parts.

2.1 Chinese Subtask

In order to make full use of the weakly supervised information calculated by the click models in the SogouQCL[17] dataset, we use Bert to predict click model scores and manual annotated score in Chinese subtask at the same time. In order to maintain the consistency of the weakly supervised signal during the training process, we only consider one of the click models PSCM, which is the closest to manual annotated score in the training dataset.

Specifically, given the training dataset $T = \{(Q_i, D_i, p_i, l_i)\}_{i=1}^{|T|}$, where p is the score calculated by PSCM click model and l is manual annotated score. $Q = \{w_1^Q, ..., w_{|Q|}^Q\}$ indicates a query and $D = \{w_1^D, ..., w_{|D|}^D\}$ indicates a document, and w is an identifier of a word. We concatenate the words in query and document with a separator *[SEP]* and add a classification head *[CLS]* in the front of the sequence. The sequence is input into Bert with 12 bidirectional Transformer [12] layers to obtain the representation of each word and separators. The model structure is shown in Fig. 1.

To predict the PSCM score and the manual annotated score at the same time, we use the representations of *[CLS]* calculated through the last two transformer layers for prediction. To be specific, the representation of *[CLS]* in the 11th layer $\boldsymbol{r}_{11}^{cls}$ is utilized to predict PSCM score p and the representation in the last layer $\boldsymbol{r}_{12}^{cls}$ is utilized to predict score human annotated score l:

$$\begin{aligned} \hat{p} &= \text{Sigmoid}\left(\boldsymbol{W}_p \boldsymbol{r}_{11}^{cls} + b_p\right), \\ \hat{l} &= \text{Sigmoid}\left(\boldsymbol{W}_l \boldsymbol{r}_{12}^{cls} + b_l\right). \end{aligned} \tag{1}$$

Where $\boldsymbol{W}_p$, $\boldsymbol{W}_l$, b_p, and b_l are parameters to be trained. Compared with the method that only uses the manually annotated label as the supervision signal, this method allows the model converging towards the ground truth through the weak supervision signal in advance. In the process of training, we use mean squared error

[1] https://github.com/microsoft/MSMARCO-Passage-Ranking

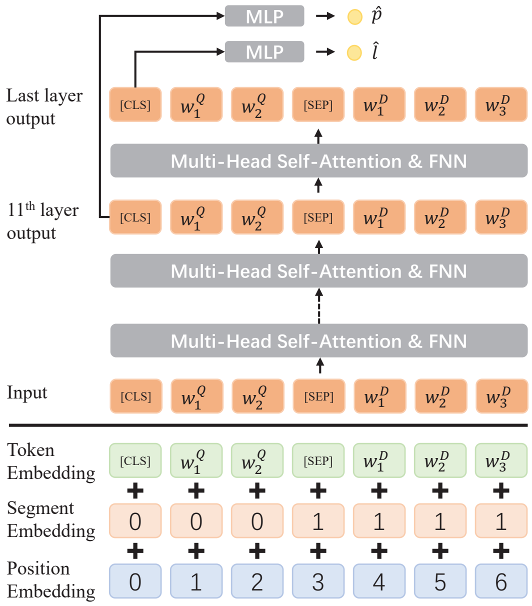

Figure 1: Structure of the model to predict PSCM score and manual annotated score based on Bert.

(MSE) as loss function:

$$\mathcal{L} = \sum_{i}^{N} (\hat{p}_i - p_i)^2 + (\hat{l}_i - l_i)^2. \tag{2}$$

2.2 English Subtask

Limited by the number of queries in the TREC Web Track dataset, it's difficult to train deep neural model on this dataset. Although the MS MARCO Passage Ranking dataset can provide a larger amount of queries, the structure of the text is different between passages and web pages. As a result, we use MS MARCO dataset to pre-train a Bert model, and fine-tune the model on TREC Web Track dataset.

The usage of Bert is similar to that in Chinese subtask, but we don't adopt PSCM score in the process of training. So we only use the representation of symbol *[CLS]* in the last bidirectional Transformer layer to predict the manual annotated score.

Additionally, we also use LambdaMART [13] to rank candidate documents for each query. We designed 41 features for learning to rank algorithm. Specifically, 24 statistic features related to length of documents, term frequency in documents, inverse document frequency and exact matching features between queries and documents. 10 semantic similarity features calculated by six types of vector distances between the representations of queries and documents. The vector distances includes Euclidean distance, Manhattan distance, cosine distance, Canberra distance and Minkowski distance. The representations of queries and documents are obtained through the sum of the Word2vec [7] embedding of each word. The last 7 features are from the results of Bert and the following 6 neural ranking models:

ARC-I [3]. A representation-focused matching model using 1d-convolution on text representation matrices.

ARC-II [3]. A interaction-focused matching model using 1d-convolution to learn the interaction representation of queries and documents, and employing 2d-convolution on the interaction matrices to obtain the ranking scores.

DSSM [4]. A representation-focused matching model by maximizing the conditional likelihood of the clicked documents given a query using the click-through data.

KNRM [15]. A interaction-focused ranking model that extracts soft-matching features between queries and documents through kernel pooling.

Duet [8]. A neural ranking model jointly utilize the local representations and distributed representations of queries and documents.

Matchpyramid [9]. A interaction-based matching model employing convolutional neural network on similarity matrices of queries and documents.

3 RUNS AND EVALUATION

3.1 Dataset

In Chinese subtask, we use Sogou-QCL and a commercial dataset. Sogou-QCL dataset contains more than 500 thousand queries and over 12 million query-document pairs. Documents in Sogou-QCL are from SogouT-16 [6] dataset, which contains about 1.17B Web pages. Actually, we only index part of the "Category B" version from SogouT-16. For each query-document pair in Sogou-QCL, 5 kinds of weak relevance labels are provided based on different click models. In our experiments, we use 2000 queries and about 50 thousand documents annotated by human to train our models. For multitask learning, we choose PSCM score as weak relevance label. Besides Sogou-QCL dataset, we also use a commercial dataset containing query log collected from a commercial search engine between 1st Jan. 2013 and 28th Feb. 2013. There are more than 188 thousand queries in the dataset, and each query contains the top 20 retrieved urls and click labels.

In English subtask, we adopt TREC Web Track dataset from 2009 to 2014. The dataset contains 300 queries and their relevance judgement files. Documents in these files are from ClueWeb09 and ClueWeb12. In addition, we also use MS MARCO Passage Ranking dataset as our training data. It contains more than 500 thousands query-passage pair and about 9 million passages.

3.2 Submitted Runs

We submit the following five runs in Chinese subtask:

RUCIR-C-CD-NEW-1. We use multi-task method described in Section 2.2 to train Bert on the commercial dataset.

RUCIR-C-CD-NEW-2. We use multi-task method described in Section 2.2 to train Bert on Sogou-QCL dataset.

RUCIR-C-CD-NEW-3. We only use manual annotation to train Bert without PSCM score on the commercial dataset.

RUCIR-C-CD-NEW-4. We only use manual annotation to train Bert without PSCM score on Sogou-QCL dataset.

Table 1: Official results of Chinese subtask

Runs	Mean nDCG	Mean Q	Mean ERR	Mean iRBU
RUCIR-C-CD-NEW-1	0.4923	0.4510	0.6029	0.8299
RUCIR-C-CD-NEW-2	0.4314	0.3887	0.5412	0.8245
RUCIR-C-CD-NEW-3	0.5136	0.4700	0.6200	0.8621
RUCIR-C-CD-NEW-4	**0.5296**	**0.4787**	**0.6442**	**0.8798**
RUCIR-C-CO-NEW-5	0.4543	0.4094	0.5456	0.8525

Table 2: Official results of Chinese subtask

Runs	Mean nDCG	Mean Q
RUCIR-E-CO-NEW-1	0.5158	0.5276
RUCIR-E-CO-NEW-2	0.5418	0.5594
RUCIR-E-CO-NEW-3	0.5363	0.5569
RUCIR-E-CO-NEW-4	0.4251	0.4207
RUCIR-E-CO-NEW-5	**0.5611**	**0.5755**

RUCIR-C-CO-NEW-5. We use multi-task method described in Section 2.2 to train Bert on Sogou-QCL dataset, but we don't use query description in the test dataset.

For English subtask, we sumbit the following five runs:

RUCIR-E-CO-NEW-1. We use MS MARCO dataset to pre-train Bert model, and use TREC dataset to fine-tune the model.

RUCIR-E-CO-NEW-2. We use MS MARCO dataset to train Bert model without fine-tuning.

RUCIR-E-CO-NEW-3. We use TREC dataset to train Bert model without fine-tuning. We use KNRM model on MS MARCO and fine-tune the model with TREC dataset.

RUCIR-E-CO-NEW-4. We use KNRM model on MS MARCO and fine-tune the model with TREC dataset.

RUCIR-E-CO-NEW-5. We use MS MARCO dataset to train the 7 neural ranking models, and use these models to collect features on TREC dataset. Then we adopt LambdaMART algorithm on TREC dataset.

3.3 Experimental Results

Official results [11] are shown in Table 1 and Table 2. We find *RUCIR-C-CD-NEW-4* achieves the best performance in Chinese subtask. Compared with *RUCIR-C-CD-NEW-2*, we find the addition of PSCM scores actually reduces the effect of the model. One possible reason is that the PSCM score is not reliable enough as weakly supervised information, or the multi-task learning method we used is not effective enough. Actually, we have tried some other methods to perform muti-task learning, such as 1) utilizing the representation of *[CLS]* in the last layer to predict PSCM score and human annotated score through two different MLP layers; 2) utilizing the representation of *[CLS]* in the previous layers of Bert to predict PSCM score. However, experiments show the model shown in Fig.1 achieves the best performance on validation set, which consists of NTCIR-13 WWW Chinese test set [5]. Additionally, The results of *RUCIR-C-CD-NEW-4* and *RUCIR-C-CD-NEW-2* are better than those of *RUCIR-C-CD-NEW-3* and *RUCIR-C-CD-NEW-1*, that may because Sogou-QCL dataset has manual annotated label, while the commercial dataset only has click results from users as relevance label.

In English subtask, *RUCIR-E-CO-NEW-5* achieves the best performance. Due to the limitation of the number of query in the training set, deep neural networks can't be trained well, so other statistic features should be added to achieve a better results.

4 CONCLUSIONS

The RUCIR team participated in the Chinese and English subtasks of the NTCIR-15 We Want Web-3 (WWW-3) Task. We applied Bert for both Chinese and English subtasks, and we used learning to rank algorithm in English subtask. We will make further explorations on how to train a deep neural network better on a limited dataset, just as we experienced on the English subtask.

ACKNOWLEDGEMENTS

Zhicheng Dou is the corresponding author. This work was supported by National Natural Science Foundation of China No. 61872370 and No. 61832017, and Beijing Outstanding Young Scientist Program NO. BJJWZYJH012019100020098.

REFERENCES

[1] Christopher J. C. Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Gregory N. Hullender. 2005. Learning to rank using gradient descent. In *Machine Learning, Proceedings of the Twenty-Second International Conference (ICML 2005), Bonn, Germany, August 7-11, 2005.* 89–96. https://doi.org/10.1145/1102351.1102363

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018). arXiv:1810.04805 http://arxiv.org/abs/1810.04805

[3] Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. 2015. Convolutional Neural Network Architectures for Matching Natural Language Sentences. *CoRR* abs/1503.03244 (2015). arXiv:1503.03244 http://arxiv.org/abs/1503.03244

[4] Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry P. Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *22nd ACM International Conference on Information and Knowledge Management, CIKM'13, San Francisco, CA, USA, October 27 - November 1, 2013.* 2333–2338. https://doi.org/10.1145/2505515.2505665

[5] Cheng Luo, Tetsuya Sakai, Yiqun Liu, Zhicheng Dou, Chenyan Xiong, and Jingfang Xu. 2017. Overview of the NTCIR-13 We Want Web Task. In *Proceedings of NTCIR-13.* to appear.

[6] Cheng Luo, Yukun Zheng, Yiqun Liu, Xiaochuan Wang, Jingfang Xu, Min Zhang, and Shaoping Ma. 2017. SogouT-16: A New Web Corpus to Embrace IR Research. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, Noriko Kando, Tetsuya Sakai, Hideo Joho, Hang Li, Arjen P. de Vries, and Ryen W. White (Eds.). ACM, 1233–1236. https://doi.org/10.1145/3077136.3080694

[7] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1301.3781

[8] Bhaskar Mitra, Fernando Diaz, and Nick Craswell. 2017. Learning to Match using Local and Distributed Representations of Text for Web Search. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*. 1291–1299. https://doi.org/10.1145/3038912.3052579

[9] Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, Shengxian Wan, and Xueqi Cheng. 2016. Text Matching as Image Recognition. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*. 2793–2799. http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/11895

[10] Stephen E. Robertson and Steve Walker. 1994. Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. In *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. Dublin, Ireland, 3-6 July 1994 (Special Issue of the SIGIR Forum)*. 232–241. https://doi.org/10.1007/978-1-4471-2099-5_24

[11] Tetsuya Sakai, Sijie Tao, Zhaohao Zeng, Yukun Zheng, Jiaxin Mao, Zhumin Chu, Yiqun Liu, Maria Maistro, Zhicheng Dou, Nicola Ferro, and Ian Soboroff. 2020. Overview of the NTCIR-15 We Want Web with CENTRE (WWW-3) Task. In *Proceedings of NTCIR-15*. to appear.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5998–6008. http://papers.nips.cc/paper/7181-attention-is-all-you-need

[13] Chao Wang, Yiqun Liu, Meng Wang, Ke Zhou, Jian-Yun Nie, and Shaoping Ma. 2015. Incorporating Non-sequential Behavior into Click Models. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9-13, 2015*, Ricardo Baeza-Yates, Mounia Lalmas, Alistair Moffat, and Berthier A. Ribeiro-Neto (Eds.). ACM, 283–292. https://doi.org/10.1145/2766462.2767712

[14] Qiang Wu, Christopher J. C. Burges, Krysta M. Svore, and Jianfeng Gao. 2010. Adapting boosting for information retrieval measures. *Inf. Retr.* 13, 3 (2010), 254–270. https://doi.org/10.1007/s10791-009-9112-1

[15] Chenyan Xiong, Zhuyun Dai, Jamie Callan, Zhiyuan Liu, and Russell Power. 2017. End-to-End Neural Ad-hoc Ranking with Kernel Pooling. *CoRR* abs/1706.06613 (2017). arXiv:1706.06613 http://arxiv.org/abs/1706.06613

[16] Jun Xu and Hang Li. 2007. AdaRank: a boosting algorithm for information retrieval. In *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007*. 391–398. https://doi.org/10.1145/1277741.1277809

[17] Yukun Zheng, Zhen Fan, Yiqun Liu, Cheng Luo, Min Zhang, and Shaoping Ma. 2018. Sogou-QCL: A New Dataset with Click Relevance Label. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, Kevyn Collins-Thompson, Qiaozhu Mei, Brian D. Davison, Yiqun Liu, and Emine Yilmaz (Eds.). ACM, 1117–1120. https://doi.org/10.1145/3209978.3210092