RUCIR at the NTCIR-16 Session Search (SS) Task

Haonan Chen
Renmin University of China
China
hnchen@ruc.edu.cn

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

ABSTRACT

This paper presents the participation of the RUCIR group, an information retrieval research group in Renmin University of China, in the NTCIR-16 Session Search Task. We will discuss the approach we used to generate the session search runs, and introduce the main experimental results. More specifically, we use the state-of-the-art context-aware ranking model COCA, which is based on BERT and contrastive learning, to generate the runs we submitted to the task. In addition, we use the BM25 algorithm and usefulness labels to make our ranking results more accurate. Official results show that our best run outperforms all other participants' runs in terms of all official metrics in both subtasks.

KEYWORDS

Session Search, BERT, Contrastive Learning, BM25

TEAM NAME

RUCIR

SUBTASKS

Fully Observed Session Search subtask (Chinese)
Partially Observed Session Search subtask (Chinese)

1 INTRODUCTION

Users' information need has become increasingly complex in recent years. In order to complete a complex search task, a user often has to issue multiple queries to find the information they need. These search behaviors that happen in a relatively short time interval are called a *search session* [12, 21]. Utilizing the contextual information of the search session has been proved beneficial for understanding the user's current search intent [1–3, 17, 23]. For example, if a user issues a query "apple" and she issues a query called "Amazon Computer for Work" a minute ago, then we can infer that she is looking for computers from the company Apple rather than the fruit apple.

The NTCIR-16 Session Search (SS) task is a new NTCIR-16 pilot task that supports comprehensive explorations of session search or task-oriented search [7]. Despite that former session-based evaluation tasks like TREC Session Tracks and Dynamic Domain (DD) Tracks have no longer been held for many years, designing and evaluating a session-based search system remains a challenge.

Most submissions of TREC 2014 Session Track [5], which is the most recent related task, used traditional retrieval approaches. For example, they used SVMrank [11], language model with Dirichlet smoothing [24], Query Change Retrieval (QCM) model [10], etc. Most of these traditional methods only focus on a few key features, which would inevitably ignore some valuable features. Many neural context-aware ranking models have been studied these years [1, 2, 6, 17, 26, 27]. Some of them [17, 26] are based on the powerful pre-trained language model BERT [9].

We implement the state-of-the-art session search model COCA [26] and use this approach for both FOSS and POSS subtasks. COCA uses three data augmentation strategies to generate similar sequences for each session sequence. Then, it utilizes a contrastive learning objective to pre-train the BERT encoder to get a more robust representation. We also apply some pre-processing techniques, e.g., identifying actual user clicks by session-level usefulness labels. In addition, we use the BM25 algorithm [18] to regularize the ranking results of COCA.

The rest of the paper is organized as follows. Related works of this article, including traditional session search models and neural approaches, are briefly introduced in Section 2. Our approach to solving the problem is introduced in Section 3. In Section 4, we describe the experimental settings and analyze the results. Finally, we make a conclusion of the whole work in Section 5.

2 RELATED WORK

2.1 Traditional models

There are already some early attempts at modeling contextual information [3, 4, 19, 20, 23]. The feasibility of lexical query matching in session search was investigated by Van Gysel et al. [20]. On TREC session query logs [1], they looked at existing lexical matching session search models. They pointed out that lexical matching techniques, particularly query term re-weighting, have a lot of room for improvement. These traditional models have demonstrated the importance and viability of session search. Due to the lack of a large-scale search log, these models are limited to static-based techniques, which are unable to comprehend the user's actual information need thoroughly.

2.2 Neural models

Ahmad et al. [1] proposed a multi-task framework that can optimize context-aware document ranking and query suggestion simultaneously. They encoded the historical queries, the current query, and the candidate document into hidden representations with several LSTMs. With these representations ready, they obtained the ranking score. Ahmad et al. [2] proposed CARS to improve their previous work. They took historical user clicks into consideration and applied the attention mechanism to get better representations.

Chen et al. [6] attempted to integrate representation and interaction for session search. Instead of matching every two behaviors in the session, they used the encoded session history to enhance the current query and the candidate document at the word-level, then matched these two by several Conv-KNRM [8] components. The results demonstrated their model's effectiveness and efficiency. Zuo

[1] http://ir.cis.udel.edu/sessions/

et al. [27] tried to model multi-granularity historical query change to identify the user's current information need. They modeled the query change on both term and semantic levels. They calculated three types of term-level weights: retained term weights, added term weights, and removed term weights. Besides, they used Transformers to combine the representations of each historical query derived using several forms of word weights, resulting in a comprehensive picture of user intent. Finally, they calculated the ranking score by combining the context-aware ranking scores acquired from term-based interaction and representation-based matching illustrated above.

The large-scale pre-trained language models like BERT [9] have achieved great success in many tasks [13–16]. These models have also been widely used in session search [17, 26]. Qu et al. [17] concatenated all user behaviors of a session (queries, clicked documents, skipped documents, and the candidate document) and put this sequence into BERT to get an overall interaction-based representation from the '[CLS]' token. Besides, they develop a hierarchical behavior-aware attention module over the BERT encoder to model interactions at different levels. Zhu et al. [26] believed that the user behavior sequence of a search session is not definite, but rather flexible. For example, a user's search behavior can vary from one search engine to another, a user can issue different queries regarding the same information need, and a user can decide whether to click a document within the same session context. Based on this idea, they explicitly generated possible variations of user behavior sequences from the search log. Based on the augmented data, they utilized contrastive learning to train the BERT model to distinguish different sequences and pull together similar ones. After this, the further pre-trained BERT encoder managed to give the session sequence a more robust representation. This is the state-of-the-art model and we implement this model to solve the problems of both subtasks.

3 METHODS

In this section, we first define the problem and some notations. Then we describe how we pre-process the data. We also illustrate COCA and how we use it in this task.

3.1 Problem Definition

Before shedding light on our solution to this problem, we want to first describe some important notations. The session context can be denoted as:

$$\mathcal{S} = \{(q_1, d_1), (q_2, d_2), ..., (q_{n-1}, d_{n-1})\}, \tag{1}$$

where q_i is the i-th query of the session and d_i is the corresponding clicked document. We also denote q_n as the current query, and d_n as the candidate document that are being scored. Note that for convenience, we will refer to q and d as the current query and the corresponding document to be ranked.

The goal of the FOSS task is to score the candidate document d under the full session context $\mathcal{S}$. For the POSS task, we simply mark the historical queries that don't have clicked documents as q_i and "[empty_d]". By this, we can still utilize the contextual information of the query sequence without being affected by some random candidate documents. Since we don't provide these two subtasks with very distinct solutions, we will focus on the FOSS task in the rest of the paper.

3.2 Pre-processing

We first extract the sessions that have human-labeled documents from the training data as the validation set and leave the rest sessions as the training set.

When building the test set, we first walk through the entire document collection to build two dictionaries. One is from a document id to its full content (did2content). The other one is from a document id to its title (did2title). We will use did2content for retrieval and did2title for re-ranking (because BERT can only take up to 512 tokens). Since we don't have the HTML or the URL of a document, we can only construct its title from its body content. We consider the first six words of its full content as its title. Besides, there is also some information of document titles in the given training and validation set which we believe is more accurate. Thus, we further use this information to supplement the did2title dictionary. The organizer also provides a dictionary that maps each query that needs evaluation to a set of candidate documents' ids (qid2docs). However, there are also some queries that are not in the given dictionary. For these queries, we use the BM25 algorithm and the did2content dictionary to retrieve fifty candidate documents for each of them and add the mapping information into qid2docs. With these data ready, we can finally build the test sets.

For each session in the test set (*e.g.*, FOSS), it has a unique session id and consists of several queries. For each historical query, it has a text, a unique query id, and ten candidate documents. Each document is comprised of a unique document id, its title, the label that indicates whether this document is clicked, the timestamp when it is clicked (-1 if not clicked), and a usefulness label that indicates its usefulness to this session judged by humans. For each query that is for evaluation, it has a text, a unique query id, and about fifty candidate documents. Each document is comprised of a unique document id and its title.

For efficiency, we only keep one clicked document per historical query. We attempt two ways to decide which document to keep: (1) We keep the document that is marked as clicked by the user in the actual search log; (2) We keep the document that has the largest usefulness value which indicates that it is the most useful document among the candidates according to humans. We will discuss these two approaches by experiments in Section 4.

3.3 COCA

The key idea of COCA [26] is using contrastive learning to optimize sequence representation of BERT encoder before ranking documents, *i.e.*, during pre-training. It has two stages: (1) pre-training stage and (2) ranking stage.

3.3.1 *Pre-training.*

COCA uses the contrastive learning technique to pull closer similar sequences and push away different ones. Specifically, as illustrated in Figure 1, it uses three data augmentation strategies to generate similar sequences:

- **Term Mask**. COCA uses a random term mask operation on the queries and documents of the session sequence. The produced sequences only change from the original one slightly.

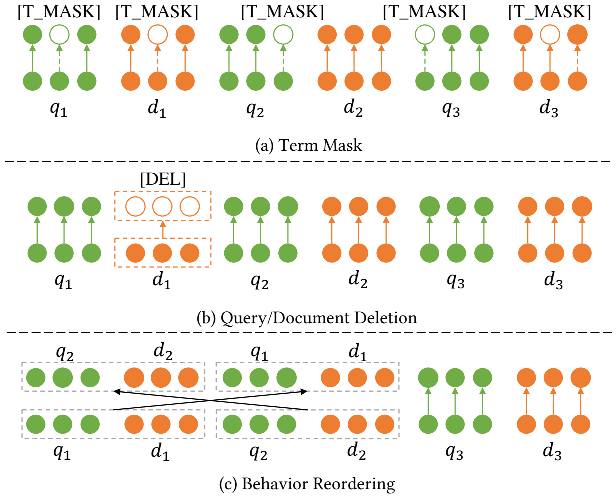

Figure 1: The augmentation strategies used in COCA. This figure is from its original paper [26].

- As a result, they are considered as similar sequences to the original one.
- **Query/Document Deletion**. COCA randomly crops the given sequence by deleting a behavior (a query or a document).
- **Behavior Reordering**. The authors of COCA believe that the order of user behavior sequence is not strict but rather flexible. COCA switches the place of two random query-document pairs to generate a similar sequence.

For all sequences in a mini-batch, COCA randomly applies two augmentation strategies to generate additional user behavior sequences. Two generated sequences from the same sequence are considered as a positive pair while the rest sequences of the mini-batch are the negative samples. COCA applies a contrastive learning objective to pull close the representations of positive pairs and push away the representations of the negative pairs. After this, the BERT encoder is able to give the user behavior sequence a more robust representation.

3.3.2 Ranking. With the further pre-trained BERT encoder ready, COCA concatenates the session sequence and put it into BERT to get a robust representation:

$$X = [\mathrm{CLS}]q_1[\mathrm{EOS}]d_1[\mathrm{EOS}]\cdots q[\mathrm{EOS}][\mathrm{SEP}]d[\mathrm{EOS}][\mathrm{SEP}].$$

$$\mathbf{r} = \mathrm{BERT}(X)_{[\mathrm{CLS}]}, \tag{2}$$

where $\mathbf{r}$ is the encoded representation of session sequence X. Then COCA gets the ranking score of d with a multi-layer perceptron (MLP):

$$P_{\mathrm{COCA}} = \mathrm{MLP}(\mathbf{r}), \tag{3}$$

For training, COCA applies a standard cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log p_i + (1 - y_i)\log(1 - p_i), \tag{4}$$

Table 1: The official results of the submitted runs of both subtasks. The best performance is in bold.

FOSS	NDCG@3	NDCG@5	NDCG@10
COCA+BM25	0.4783	0.4785	0.4939
COCA+U	0.5365	0.5406	0.5570
COCA+BM25+U	**0.5525**	**0.5623**	**0.5693**
POSS	**RsDCG**	**RsRBP**	
COCA	0.4355	0.5640	
COCA+BM25	0.4738	0.6281	
COCA+BM25+U	**0.5439**	**0.7466**	

where N is the number of the samples in the training set and y_i is the label.

3.3.3 With BM25. BM25 [18] is a traditional retrieval algorithm that ranks the candidate documents based on the terms of the current query appearing in them. In order to prevent overfitting, we use the BM25 score of the documents to regularize the ranking score given by COCA as follows:

$$P(d|S, q) = \alpha P_{\mathrm{COCA}} + (1 - \alpha)P_{\mathrm{BM25}}, \tag{5}$$

where α is a hyper-parameter that is tuned on the validation set.

4 EXPERIMENTS

4.1 Metrics

The official metric of the FOSS subtask is Normalized Discounted Cumulative Gain (NDCG) [22]. The metrics of POSS are RsDCG and RsRBP [25].

4.2 Implementation Detail

We select the hyperparameters of COCA following its original paper [26]. The value of α is tuned on the validation set to be 0.67.

We implement several variants of COCA: (1) +U: The model keeps the document that has the largest usefulness value as illustrated in Section 3.2. (2) +BM25: The model has BM25 as regularization.

4.3 Results and Analysis

The official results of our submitted runs of both subtasks are shown in Table 1. The model with BM25 as regularization and taking usefulness labels into consideration achieves the highest results. For example, it achieves about 0.57 in terms of NDCG@10 in the FOSS subtask. We can draw the following conclusions from the experimental results:

(1) Using BM25 as regularization may benefit the performance of neural ranking models. We can observe that COCA performs better with the BM25 score as regularization in both subtasks. This indicates that traditional retrieval algorithms like BM25 may help neural models prevent overfitting.

(2) Usefulness labels annotated by humans can reduce the noise of session history. It is obvious that COCA performs much better with the usefulness labels being considered. We believe that the documents that are marked clicked are not always relevant to the session because users may click some irrelevant documents just to explore their interests or feel lucky. However, usefulness labels annotated by humans are far more reliable to be treated as relevance to the session.

5 CONCLUSIONS

This paper presents our participation in the Session Search task at NTCIR-16. Our model is based on the state-of-the-art session search model COCA. We also utilize the BM25 algorithm to prevent overfitting and the usefulness labels to denoise the session context. Our best run achieves the top performance in both subtasks in terms of all official metrics.

ACKNOWLEDGMENTS

Zhicheng Dou is the corresponding author. This work was supported by the National Natural Science Foundation of China No. 61872370.

REFERENCES

[1] Wasi Uddin Ahmad, Kai-Wei Chang, and Hongning Wang. [n.d.]. Multi-Task Learning for Document Ranking and Query Suggestion. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings.*

[2] Wasi Uddin Ahmad, Kai-Wei Chang, and Hongning Wang. 2019. Context Attentive Document Ranking and Query Suggestion. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019.*

[3] Paul N. Bennett, Ryen W. White, Wei Chu, Susan T. Dumais, Peter Bailey, Fedor Borisyuk, and Xiaoyuan Cui. [n.d.]. Modeling the impact of short- and long-term behavior on search personalization. *SIGIR'12 - Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval* ([n. d.]).

[4] Ben Carterette, Paul Clough, Mark Hall, Evangelos Kanoulas, and Mark Sanderson. 2016. Evaluating retrieval over sessions: The TREC session track 2011-2014. *SIGIR 2016 - Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2016).

[5] Ben Carterette, Evangelos Kanoulas, Mark Hall, and Paul Clough. 2014. *Overview of the TREC 2014 session track.* Technical Report.

[6] Haonan Chen, Zhicheng Dou*, Qiannan Zhu, Xiaochen Zuo, and Ji-Rong Wen. 2022. Integrating Representation and Interaction for Context-Aware Document Ranking. *ACM Trans. Inf. Syst.* (2022). https://doi.org/10.1145/3529955

[7] Jia Chen, Weihao Wu, Jiaxin Mao, Beining Wang, Fan Zhang, and Yiqun Liu. 2022. Overview of the NTCIR-16 Session Search (SS) Task. *Proceedings of NTCIR-16. to appear* (2022).

[8] Zhuyun Dai, Chenyan Xiong, Jamie Callan, and Zhiyuan Liu. 2018. Convolutional Neural Networks for Soft-Matching N-Grams in Ad-hoc Search. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, Marina Del Rey, CA, USA, February 5-9, 2018.* ACM, 126–134. https://doi.org/10.1145/3159652.3159659

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers).*

[10] Dongyi Guan, Sicong Zhang, and Hui Yang. 2013. Utilizing query change for session search. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval.*

[11] Thorsten Joachims. 2006. Training linear SVMs in linear time. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining.*

[12] Rosie Jones and Kristina Lisa Klinkner. 2008. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008, Napa Valley, California, USA, October 26-30, 2008.*

[13] Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UnifiedQA: Crossing Format Boundaries With a Single QA System. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020, Online Event, 16-20 November 2020.*

[14] Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020.*

[15] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-Task Deep Neural Networks for Natural Language Understanding. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers.*

[16] Xinyu Ma, Jiafeng Guo, Ruqing Zhang, Yixing Fan, Xiang Ji, and Xueqi Cheng. 2021. PROP: Pre-training with Representative Words Prediction for Ad-hoc Retrieval. In *WSDM '21, The Fourteenth ACM International Conference on Web Search and Data Mining, Virtual Event, Israel, March 8-12, 2021.*

[17] Chen Qu, Chenyan Xiong, Yizhe Zhang, Corby Rosset, W. Bruce Croft, and Paul Bennett. 2020. Contextual Re-Ranking with Behavior Aware Transformers. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020.*

[18] Stephen E. Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.* (2009).

[19] Xuehua Shen, Bin Tan, and Chengxiang Zhai. 2005. Context-sensitive information retrieval using implicit feedback. *SIGIR 2005 - Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (2005).

[20] Christophe Van Gysel, Evangelos Kanoulas, and Maarten De Rijke. 2016. Lexical query modeling in session search. *ICTIR 2016 - Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval* (2016).

[21] Hongning Wang, Yang Song, Ming Wei Chang, Xiaodong He, Ryen W. White, and Wei Chu. 2013. Learning to extract cross-session search tasks. *WWW 2013 - Proceedings of the 22nd International Conference on World Wide Web* (2013).

[22] Yining Wang, Liwei Wang, Yuanzhi Li, Di He, and Tie-Yan Liu. 2013. A theoretical analysis of NDCG type ranking measures. In *Conference on Learning Theory.*

[23] Ryen W. White, Wei Chu, Ahmed Hassan, Xiaodong He, Yang Song, and Hongning Wang. [n.d.]. Enhancing personalized search by mining and modeling task behavior. *WWW 2013 - Proceedings of the 22nd International Conference on World Wide Web* ([n. d.]).

[24] Chengxiang Zhai and John Lafferty. 2004. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS)* (2004).

[25] Fan Zhang, Jiaxin Mao, Yiqun Liu, Weizhi Ma, Min Zhang, and Shaoping Ma. 2020. Cascade or Recency: Constructing Better Evaluation Metrics for Session Search. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020.*

[26] Yutao Zhu, Jian-Yun Nie, Zhicheng Dou, Zhengyi Ma, Xinyu Zhang, Pan Du, Xiaochen Zuo, and Hao Jiang. 2021. Contrastive Learning of User Behavior Sequence for Context-Aware Document Ranking. In *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021.*

[27] Xiaochen Zuo, Zhicheng Dou, and Ji-Rong Wen. 2022. Improving Session Search by Modeling Multi-Granularity Historical Query Change. In *WSDM '22: The Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event / Tempe, AZ, USA, February 21 - 25, 2022.* ACM, 1534–1542. https://doi.org/10.1145/3488560.3498415