THUIR_SS at the NTCIR-17 Session Search (SS) Task

Xinyan Han[1], Yiteng Tu[2], Haitao Li[1], Qingyao Ai[1], Yiqun Liu[1]
1 Department of Computer Science and Technology, Institute for Artificial Intelligence,
Beijing National Research Center for Information Science and Technology,
Tsinghua University, Beijing 100084, China
2 Renmin University of China, Beijing 100872, China

❖ Introduction

- **We participated in FOSS and POSS subtasks in NTCIR17 Session Search task.**
  - In both subtasks, we tried different approaches for feature fusion, including Learning-to-Rank and linearly combination.
  - The final report of the SS-2 task demonstrate the effectiveness of our method, significantly outperforming other competitors.

❖ FOSS Subtask

- In FOSS subtask, we concatenate all the queries and the first clicked document title of each query except the last one in a session as session context.

Learning-to-Rank

- In our approaches, we incorporate a total of 11 features which include both term-level features from traditional sparse retrieval methods and semantic-level features obtained through deep neural networks.
- We choose two classic but effective sparse retrieval methods, BM25 and QLD.
- We train a dense retrieval model on on TianGong-ST, using the InfoNCE Loss as the loss function .
- We also train a fine-grained context-aware ranking model, DCL, with a curriculum learning framework.
- We feed the 11 features (Table 1) into two widely-used learning-to-rank models, LightGBM and LambdaMART.

	feature
1	ad-hoc score of BM25
2	ad-hoc score of QLD
3	ad-hoc score of BM25 with RM3
4	ad-hoc score of QLD with RM3
5	session score of BM25
6	session score of QLD
7	session score of BM25 with RM3
8	session score of QLD with RM3
9	ad-hoc score of DR model
10	session score of DR model
11	sesion score of DCL model

Table 1: features of learning-to-rank model

Linearly combination

- We use the linear combination of two ad-hoc scores to generate LS score.
- In THUIR_SS-FOSS-NEW-3 and THUIR_SS-POSS-NEW-3, we choose the ad-hoc score of BM25 as S1 and the score computed by DCL model as S2.
- In submission THUIR_SS-FOSS-NEW-6 and THUIR_SS-POSS-NEW-6, we replace S1 with QLD ad-hoc score.

RRF

- We sort documents according to the score of 11 features. Then we use all 11 rankings to calculate RRF scores.

❖ POSS Subtask

- In POSS subtask, the user interaction information for the last k-n queries is not provided, we just skip the clicked document and concatenate query.
- We use the same method of FOSS subtask to rank documents in POSS subtask.

❖ Submitted Runs and Evaluation

FOSS subtask.

- The preliminary evluation of our runs in FOSS subtask are shown in Table 2.
- The linear combination method using the ad-hoc score of BM25 and the score computed by DCL model achieved the best performance.

Run Name	Description	NDCG@3	NDCG@5	Rank
THUIR_SS-FOSS-NEW-1	learning-to-rank (LightGBM)	0.1547940	0.2038491	5
THUIR_SS-FOSS-NEW-3	linear combination	0.5853154	0.6745773	1
THUIR_SS-FOSS-NEW-4	learning-to-rank (LambdaMART)	0.2506041	0.3309875	4
THUIR_SS-FOSS-NEW-5	RRF	0.3931865	0.4768206	3
THUIR_SS-FOSS-NEW-6	linear combination	0.5643186	0.6569274	2

Table 2: Preliminary Evaluation of Our Runs in FOSS Subtask

POSS subtask.

- The preliminary evluation of our runs in POSS subtask are shown in Table 3.
- The linear combination method using the ad-hoc score of QLD and the score computed by DCL model achieved the best performance.

Run Name	Description	RS_DCG	RS_RBP	Rank
THUIR_SS-POSS-NEW-1	learning-to-rank (LightGBM)	0.023533	0.048312	5
THUIR_SS-POSS-NEW-3	linear combination	0.174898	0.367266	2
THUIR_SS-POSS-NEW-4	learning-to-rank (LambdaMART)	0.068510	0.143386	4
THUIR_SS-POSS-NEW-5	RRF	0.136628	0.288760	3
THUIR_SS-POSS-NEW-6	linear combination	0.181201	0.379338	1

Table 3: Preliminary Evaluation of Our Runs in POSS Subtask

- Our learning-to-rank method still has room for improvement.
  - First, more features can be selected to feed into the model, such as user interaction information (clicks and timestamp)
  - Second, we find that sometimes the query at the beginning of a session and the query at the end of a session are not very semantically related. These unrelated queries bring noise to our result.

❖ Conclusion & Future work

- Our team (THUIR_SS) participates in the FOSS and POSS subtask of the NTCIR-17 Session Search (SS) Task.
- We try learning to rank model, RRF method, and linear combination method.
- The submission using linear combination achieves the best performance in both FOSS and POSS subtasks.

Email: yitengtu16@gmail.com