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, Tsinghua University,

Zhongguancun Laboratory, Beijing 100084, China

2 Renmin University of China, Beijing 100872, China

yitengtu16@gmail.com

Introduction

- Session Search holds significant importance in the field of information retrieval and user experience.

- We participated in both FOSS and POSS subtasks in NTCIR17 Session Search task.

- Our methods significantly outperform other competitors.

Methods

Search Context Definition

FOSS: $C = [q_1, d_1^+, \ldots, q_{k-1}, d_{k-1}^+, q_k]$

POSS: $C = [q_1, d_1^+, \ldots, q_n, d_n^+, q_{n+1}, \ldots, q_j], \quad (n+1 \leq j \leq k)$

q_i is the i-th query in a session
d_i^+ is the first clicked document title of q_i (may be empty)

Methods

Feature Extraction

11 features: **both term level and semantic level, both single-turn level and session level**

• 2 classic but effective sparse retrieval methods, BM25 and QLD
• A dense retrieval model trained on TianGong-ST using contrastive learning
• A context-aware ranking model, DCL, trained with a dual curriculum learning framework

	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

Methods

Feature Extraction

Sparse Retrieval

- Model Selection: BM25 & QLD

- Retrieval Strategy: w/ & w/o RM3 pseudo-relevance feedback

- Query Text: Single-turn Query q_k & Session Query C

$$BM25(d,q) = \sum_{t \in q} \frac{IDF(t) * TF(t,d) * (k_1 + 1)}{TF(t,d) + k_1 * (1 - b + b * \frac{len(d)}{avgdl}}\,)$$

$$QLD(d,q) = \prod_{t \in q} P(t|d)^{c(t,q)}$$

Methods

Feature Extraction

Dense Retrieval

- Model Structure: Dual-encoder

$$r_q = BERT_{[CLS]}(q)$$

$$r_d = BERT_{[CLS]}(d)$$

$$s(q, d) = r_q \cdot r_d$$

- Training Data: Tiangong-ST

- Training Technique: Contrastive Learning (InfoNCE Loss)

$$\mathcal{L} = -log \frac{\exp\left(s(q, d^+)\right)}{\exp\left(s(q, d^+)\right) + \sum_j \exp\left(s(q, d_j^-)\right)}$$

Methods

Feature Extraction

Context-aware Ranking Model： DCL

- Model Structure: BERT + MLP

- Training Data: Tiangong-ST

- Training Technique: A Dual Curriculum Learning Framework*

$$X = [CLS]q_1[EOS]d_1[EOS]...q[EOS][SEP]d[EOS][SEP]$$

$$r = BERT(X)_{[CLS]}$$

$$OUTPUT = MLP(r)$$

* Zhu et al. "From Easy to Hard: A Dual Curriculum Learning Framework for Context-Aware Document Ranking." *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 2022.

Methods

Feature Extraction

Context-aware Ranking Model

- The Curriculum Learning Framework

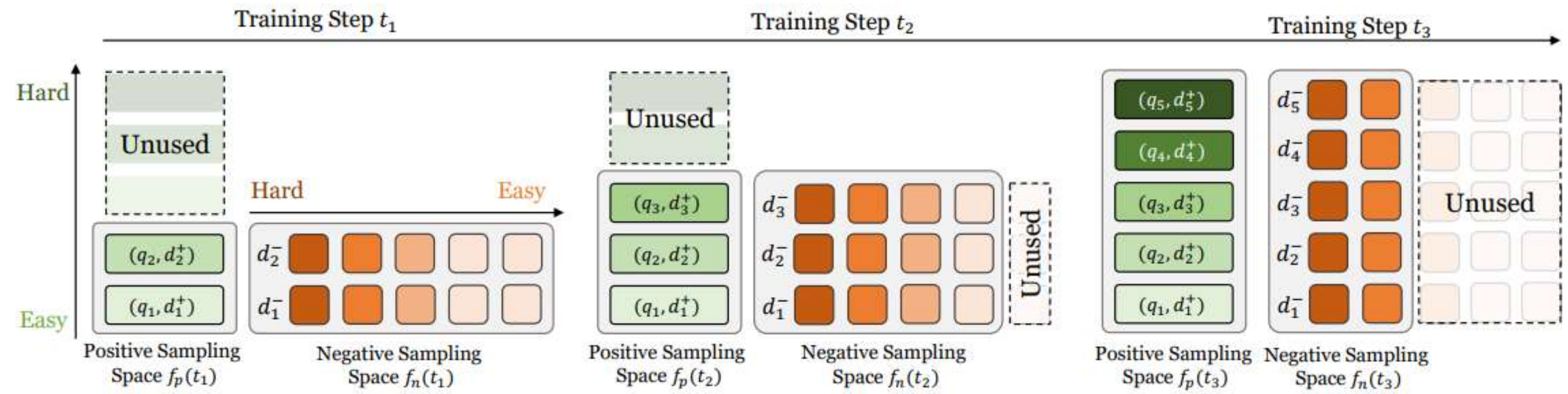

Figure 2: The training process of our framework. For the curriculum of positive pairs, only easy samples are used at the beginning (t_1). Along the training process, the positive sampling space is gradually extended to the whole positive pairs (t_3). For the curriculum of negative pairs, the sampling space is shrinking from all samples (t_1) to only hard samples (t_3).

* Zhu et al. "From Easy to Hard: A Dual Curriculum Learning Framework for Context-Aware Document Ranking." *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 2022.

Methods

Fusion

- RRF

$$RRF(d \in D) = \sum_{r \in R} \frac{1}{k + r(d)}$$

- Linear Combination
  - 1 sparse score + 1 DCL score
- Learning-to-rank
  - LightGBM
  - 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

Results and Future Work

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

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

- Linear combination achieves the best performance
- Room for improvements:
  - More features (user interaction, query/document length ……)
  - Semantic inconsistency in session

Conclusion

- Our team (THUIR_SS) participate in the FOSS and POSS subtask of the NTCIR-17 Session Search (SS) Task.
- We try several retrieval and ranking approaches on different levels and explore different fusion methods.
- The submission using linear combination achieves the best performance in both FOSS and POSS subtasks.

Thanks!

Q&A