THUIR at the NTCIR-18 FairWeb-2 Task

Huixue Su[2], Yiteng Tu[1], 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

❖ Introduction

- **We participated in the NTCIR-18 Fairweb-2 Task.**
  - For the Web Search (WS) Subtask, we utilize several different reranking models in all 5 submitted runs including classic ones like MonoBERT and MonoT5 and emerging ones like Qwen2 reranker and BGE reranker.
  - The official results indicate that our approaches achieve promising results on all relevance and fairness metrics.

❖ Our Methods

- **Retrieval Pipeline**
  - Sparse Retrieval
    - Two classic sparse retrieval algorithms, BM25 and QLD.
    - Three types of query, *Q/D/QD-queries*.
  - Neural Reranking
    - Classic: MonoBERT, MonoT5
    - New: Qwen2 Reranker, BGE Reranker
  - Result Aggregation
    - Reciprocal Rank Fusion (RRF)
- **Run 1: Qwen2-reranker-1.5B**
  - We employ Qwen2-reranker-1.5B, which leverages Qwen-1.5B as backbone and cross-encoder architecture, in order to better capture intricate semantic interactions.
  - Rerank all retrieved documents of the *QD* types of queries from sparse retrieval, 2 features in total.
- **Run 2: BGE-reranker-v2-gemma**
  - We employ BGE-reranker-v2-gemma, which is a core component of the BAAI General Embedding (BGE) series and is designated as a cross-encoder to directly assess the relevance of query-document pairs.
  - Rerank all retrieved documents of the *QD* types of queries from sparse retrieval, 2 features in total.
- **Run 3: Also BGE-reranker-v2-gemma**
  - Rerank all retrieved documents of the *Q/D/QD* types of queries from sparse retrieval, 6 features in total.
- **Run 4 & Run 5: Fusion of Reranking Results**
  - *qwen2-reranker-1.5B, bge-reranker-v2-gemma, bge-reranker-v2-m3, monobert-large-msmarco, monot5-3b-msmarco-10k*
  - Run 4: Former 3 models with *QD*, 6 features in total.
  - Run 5: All 5 models with *Q/D/QD*, 30 features in total.
- **Revived Run: PM2 & xQuAD**
  - We attempt two different ways of estimating attribute scores of each candidate document, including extracting possible entities and obtaining attribute information through web crawlers, and simply approximating attribute distribution through proportions of related term appeared in documents.
  - We try two search result diversification algorithms, PM2 and xQuAD, to balance both relevance and fairness factors.

❖ Results and Analysis

Table 1: The official relevance evaluation of our runs.

	R topics		M topics		Y topics		All topics	
Run	Mean ERR	Mean iRBU	Mean ERR	Mean iRBU	Mean ERR	Mean iRBU	Mean ERR	Mean iRBU
THUIR-WS-QD-RR-1	0.1706	0.3781	0.1483	0.4237	0.0908	0.2517	0.1380 (>22-23)	0.3523 (>22-23)
THUIR-WS-QD-RR-2	0.1661	0.4106	0.2463	0.5017	0.0864	0.2164	0.1662 (>19-23)	0.3777 (>20-23)
THUIR-WS-QD-RR-3	0.1254	0.4164	0.2209	0.5301	0.0734	0.2553	0.1393 (>22-23)	0.4012 (>19-23)
THUIR-WS-QD-RR-4	0.1682	0.4211	0.2027	0.4725	0.0944	0.2493	0.1557 (>19-23)	0.3827 (>19-23)
THUIR-WS-QD-RR-5	0.1209	0.4193	0.2216	0.5171	0.0794	0.2661	0.1398 (>22-23)	0.4016 (>19-23)
THUIR-WS-QD-REV-1	**0.2347**	0.4893	0.2367	0.5804	0.0636	0.2327	0.1807 (>18-23)	0.4365 (>17-23)
Best of Other Participants	0.2322	**0.5357**	**0.2700**	**0.6332**	**0.1191**	**0.3024**	**0.2020 (>12-23)**	**0.4613 (>15-23)**
ORG-WS-run.bm25.D	0.0444	0.2149	0.1074	0.3413	0.0382	0.1186	0.0625	0.2245 (>23)
ORG-WS-run.bm25.Q	0.0987	0.3258	0.1565	0.5300	0.0724	0.2495	0.1088 (>23)	0.3666 (>22-23)
ORG-WS-run.qld.D	0.0441	0.2061	0.0882	0.3133	0.0383	0.1171	0.0563	0.2119 (>23)
ORG-WS-run.qld.Q	0.1079	0.3196	0.1320	0.4447	0.0597	0.2104	0.1002 (>23)	0.3246 (>23)
ORG-WS-run.qljm.D	0.0335	0.1424	0.1196	0.3385	0.0482	0.1834	0.0656	0.2181 (>23)
ORG-WS-run.qljm.Q	0.1303	0.3161	0.0988	0.3027	0.1017	**0.3024**	0.1111 (>23)	0.3074 (>23)

Table 2: The official fairness evaluation over the R topics of our runs.

Run	Mean GF^{JSD} (PRONOUN)	Mean GF^{NMD} (HINDEX)	Mean GF^{RNOD} (HINDEX)	Mean GFR
THUIR-WS-QD-RR-1	0.3344 (>23)	0.3323 (>23)	0.2977 (>23)	0.3367 (>23)
THUIR-WS-QD-RR-2	0.3761 (>23)	0.3751 (>23)	0.3428 (>23)	0.3765 (>23)
THUIR-WS-QD-RR-3	0.4070 (>22-23)	0.3945 (>22-23)	0.3701 (>22-23)	0.3978 (>22-23)
THUIR-WS-QD-RR-4	0.3866 (>23)	0.3860 (>23)	0.3528 (>23)	0.3868 (>23)
THUIR-WS-QD-RR-5	0.4233 (>22-23)	0.4028 (>22-23)	0.3775 (>22-23)	0.4067 (>22-23)
THUIR-WS-QD-REV-1	0.4247 (>22-23)	0.4226 (>22-23)	0.3800 (>22-23)	0.4314 (>22-23)
Best of Other Participants	**0.4687 (>21-23)**	**0.4716 (>18-23)**	**0.4248 (>21-23)**	**0.4764 (>18-23)**
ORG-WS-run.bm25.D	0.2265	0.2092	0.1991	0.2135
ORG-WS-run.bm25.Q	0.3300 (>23)	0.3157 (>23)	0.2984 (>23)	0.3181 (>23)
ORG-WS-run.qld.D	0.2125	0.2054	0.1988	0.2058
ORG-WS-run.qld.Q	0.3143 (>23)	0.3015 (>23)	0.2777 (>23)	0.3039 (>23)
ORG-WS-run.qljm.D	0.1436	0.1413	0.1335	0.1398
ORG-WS-run.qljm.Q	0.2938 (>23)	0.2872 (>23)	0.2645 (>23)	0.2914 (>23)

Table 3: The official fairness evaluation over the M topics of our runs.

Run	Mean GF^{JSD} (ORIGIN)	Mean GF^{NMD} (RATINGS)	Mean GF^{RNOD} (RATINGS)	Mean GFR
THUIR-WS-QD-RR-1	0.3359 (>23)	0.3769 (>23)	0.3514 (>23)	0.3704 (>23)
THUIR-WS-QD-RR-2	0.3411 (>23)	0.4145 (>23)	0.3704 (>23)	0.4044 (>23)
THUIR-WS-QD-RR-3	0.4028 (>23)	0.4610 (>23)	0.4194 (>23)	0.4508 (>23)
THUIR-WS-QD-RR-4	0.3365 (>23)	0.3965 (>23)	0.3590 (>23)	0.3893 (>23)
THUIR-WS-QD-RR-5	0.3757 (>23)	0.4461 (>23)	0.4030 (>23)	0.4320 (>23)
THUIR-WS-QD-REV-1	0.4034 (>23)	0.4973 (>22-23)	0.4437 (>23)	0.4758 (>22-23)
Best of Other Participants	**0.4491 (>23)**	**0.5207 (>22-23)**	**0.4644 (>22-23)**	**0.5101 (>22-23)**
ORG-WS-run.bm25.D	0.2800	0.3192 (>23)	0.2917 (>23)	0.3043 (>23)
ORG-WS-run.bm25.Q	0.4331 (>23)	0.4948 (>22-23)	0.4509 (>23)	0.4713 (>23)
ORG-WS-run.qld.D	0.2706	0.3000	0.2784	0.2874
ORG-WS-run.qld.Q	0.3770 (>23)	0.4203 (>23)	0.3862 (>23)	0.4026 (>23)
ORG-WS-run.qljm.D	0.2644	0.3069 (>23)	0.2792	0.2940 (>23)
ORG-WS-run.qljm.Q	0.2560	0.2809	0.2587	0.2725

Table 4: The official fairness evaluation over the Y topics of our runs.

Run	Mean GF^{NMD} (SUBSCS)	Mean GF^{RNOD} (SUBSCS)	Mean GFR
THUIR-WS-QD-RR-1	0.2247	0.2153	0.2335
THUIR-WS-QD-RR-2	0.1919	0.1835	0.2000
THUIR-WS-QD-RR-3	0.2407 (>23)	0.2322 (>23)	0.2437 (>23)
THUIR-WS-QD-RR-4	0.2211	0.2109	0.2301
THUIR-WS-QD-RR-5	0.2484 (>23)	0.2401 (>23)	0.2531 (>23)
THUIR-WS-QD-REV-1	0.2179	0.2086	0.2206
Best of Other Participants	0.2109	0.2032	0.2226
ORG-WS-run.bm25.D	0.1086	0.0994	0.1090
ORG-WS-run.bm25.Q	0.2367	0.2240	0.2368
ORG-WS-run.qld.D	0.1072	0.0993	0.1082
ORG-WS-run.qld.Q	0.2003	0.1923	0.2013
ORG-WS-run.qljm.D	0.1751	0.1671	0.1753
ORG-WS-run.qljm.Q	**0.2659 (>23)**	**0.2526 (>20-23)**	**0.2775 (>21-23)**

- **Our methods achieve promising results on all relevance metrics and fairness metrics.**
  - Among all our runs through all topics, REV performs the best.
  - Comparison between Run1 and Run2 shows that BGE better boosts relevance performance and balanced entity attributes.
  - Run 4 and Run 5 form the top cluster, which likely benefits from complementary strengths across models.

❖ Conclusions

- **We participate in the NTCIR-18 FairWeb-2 task and submit 5 runs using various methods.**
- **We achieve second place in all metrics.**
- **Our results indicate that the relevance and fairness are complementary to each other to some extent. Improving relevance can also have some positive impact on fairness at the same time.**

Email: suhuixue@ruc.edu.cn