Round 2 results — Run random_bert_tiab submitted from Random

Run Description

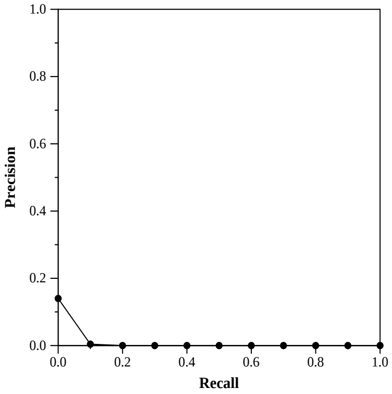
We used pre-trained BERT model and fine-tuned it with relevance judgement from round 1. We concatenated Query and Document (only title and abstract in our case, we assumed if the document is relevant to the query then the title and abstract is relevant to the query and vice-versa) for BERT input: [[CLS] Q [SEP] D [SEP]]. We used 428 tokens for BERT input, if our Query+Document Title+abstract were less than 428 tokens, we padded it with zero and if it was more than 428 tokens, we truncated it. Next, using the CORD-19 dataset provided for round 2, we retrieved 3000 documents per topic using basic tf-idf model with cosine-similarity as ranking score. Then we used our fine-tuned BERT model to classify each query-document pair (retrieved from our tf-idf model) as relevant, partially relevant or not relevant. Finally, we retrieved 1000 documents for each topic based on the classification results from our BERT model. The ranking score for our final BERT model is the classification probability score in descending order.

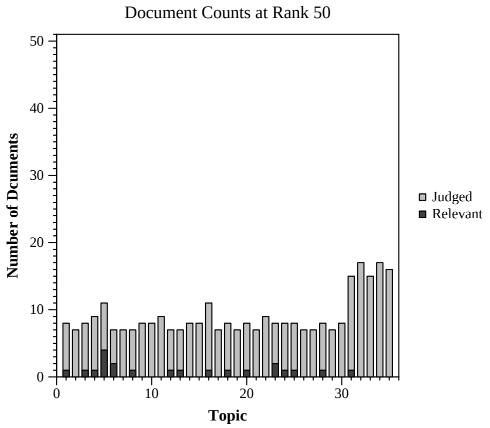
Summary Statistics	
Run ID	random_bert_tiab
Topic type	feedback
Contributed to judgment sets?	yes

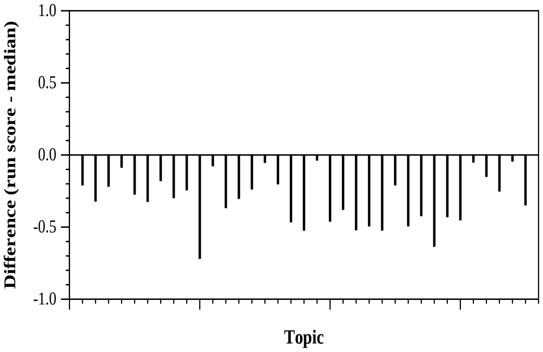
Overall measures	
Number of topics	35
Total number retrieved	33814
Total relevant	3002
Total relevant retrieved	248
MAP	0.0028
Mean Bpref	0.0583
Mean NDCG@10	0.0342
Mean RBP(p=0.5)	0.0384 +0.0066

Document Level Averages	
	Precision
At 5 docs	0.0514
At 10 docs	0.0429
At 15 docs	0.0286
At 20 docs	0.0214
At 30 docs	0.0162

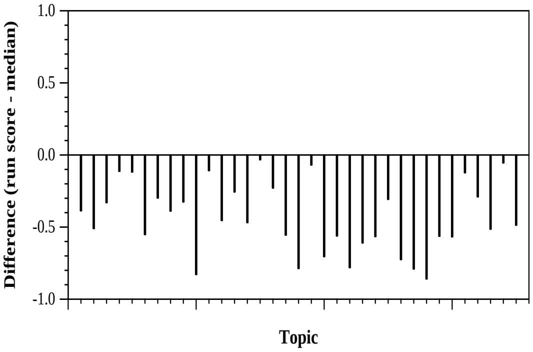
R-Precision	
Exact	0.0093





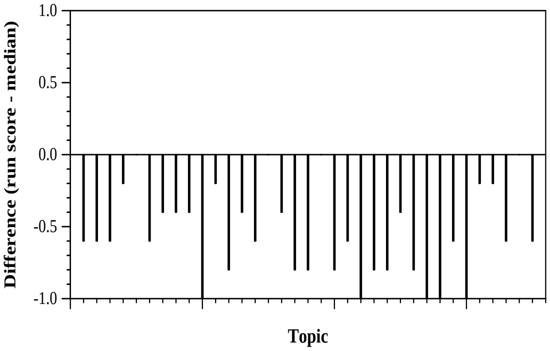


Per-topic difference from median bpref for all Round 2 runs

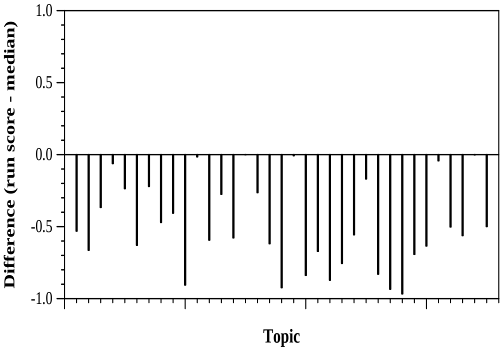


Per-topic difference from median NDCG@10 for all Round 2 runs

Round 2 results — Run random_bert_tiab submitted from Random



Per-topic difference from median P@5 for all Round 2 runs



Per-topic difference from median RBP(p=0.5) for all Round 2 runs