Round 2 results — Run factum-hybrid-qa submitted from Factum

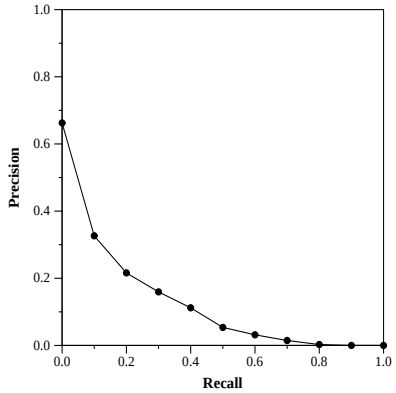
Run Description

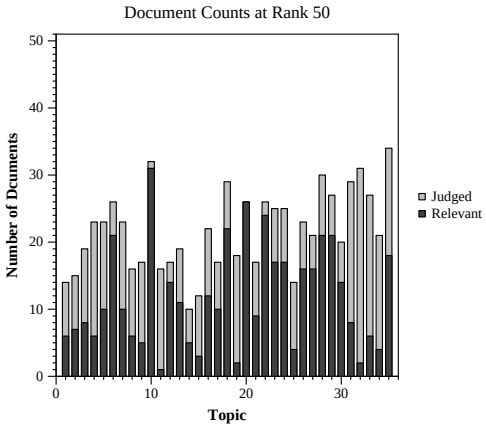
We used a siamese BERT encoder trained on NLI and QA data with sampled softmax loss to encode snippets of up to 5 sentences from the paragraphs of the articles. The final score is derived by a combination of cosine similarity of the dense vectors, BM25 score of the sparse vectors (unigrams + bigrams with lemmatisation) and the answer confidence (if found) by an extractive QA model (ALBERT-base trained on SQUAD 2.0). We removed articles published before December 2019.

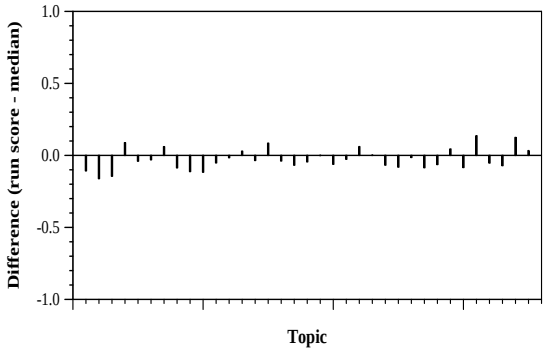
Summary Statistics	
Run ID	factum-hybrid-qa
Topic type	automatic
Contributed to judgment sets?	yes

Overall measures	
Number of topics	35
Total number retrieved	29409
Total relevant	3002
Total relevant retrieved	1471
MAP	0.1138
Mean Bpref	0.3417
Mean NDCG@10	0.3106
Mean RBP(p=0.5)	0.3290 +0.0037

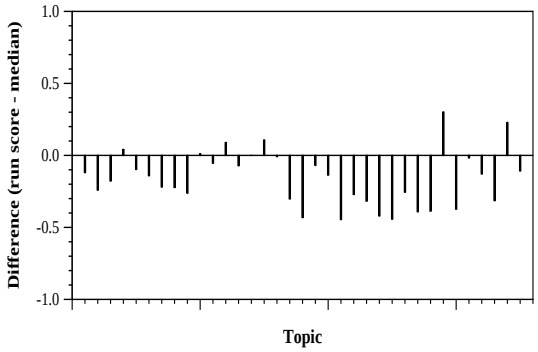
Document Level Averages	
	Precision
At 5 docs	0.4000
At 10 docs	0.3686
At 15 docs	0.3429
At 20 docs	0.3157
At 30 docs	0.2781
R-Precision	
Exact	0.1960





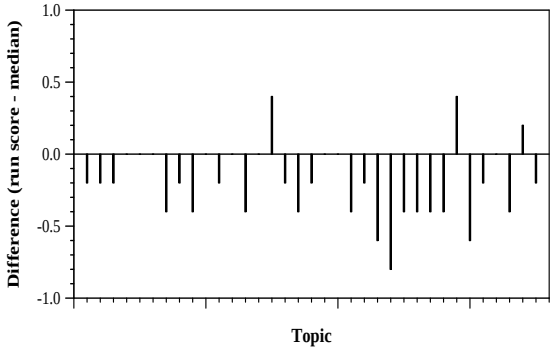


Per-topic difference from median bpref for all Round 2 runs

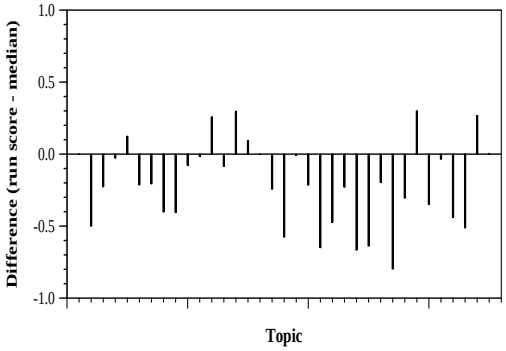


Per-topic difference from median NDCG@10 for all Round 2 runs

Round 2 results — Run factum-hybrid-qa submitted from Factum



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