

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

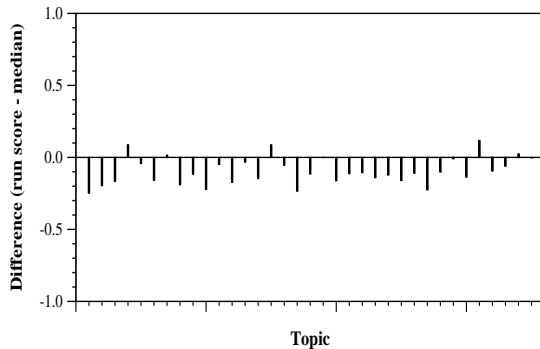
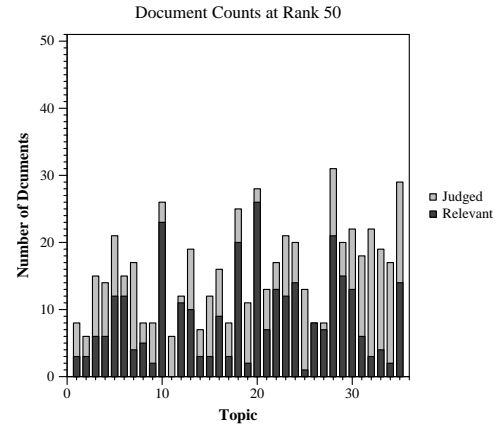
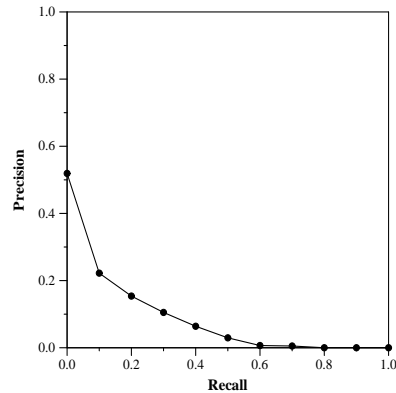
Run Description

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 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.

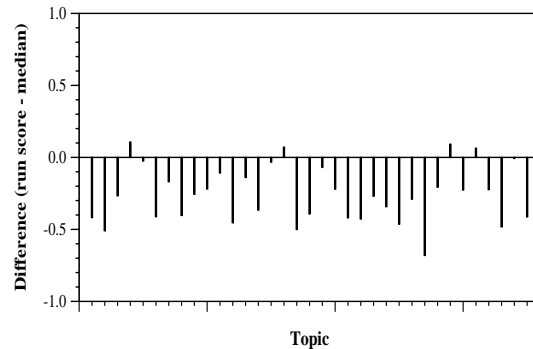
Summary Statistics	
Run ID	factum-dense-qa
Topic type	automatic
Contributed to judgment sets?	no

Overall measures	
Number of topics	35
Total number retrieved	19687
Total relevant	3002
Total relevant retrieved	1069
MAP	0.0746
Mean Bpref	0.2743
Mean NDCG@10	0.2124
Mean RBP(p=0.5)	0.2058 +0.3478

Document Level Averages	
	Precision
At 5 docs	0.2457
At 10 docs	0.2600
At 15 docs	0.2476
At 20 docs	0.2214
At 30 docs	0.1886
R-Precision	
Exact	0.1389

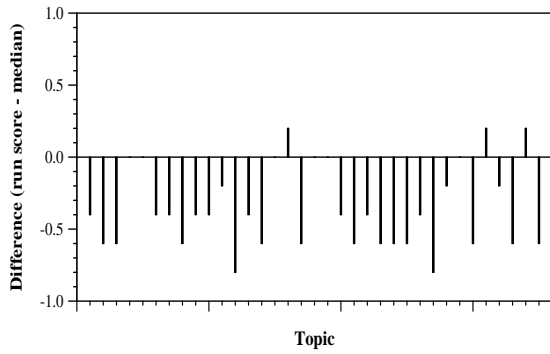


Per-topic difference from median bpref for all Round 2 runs

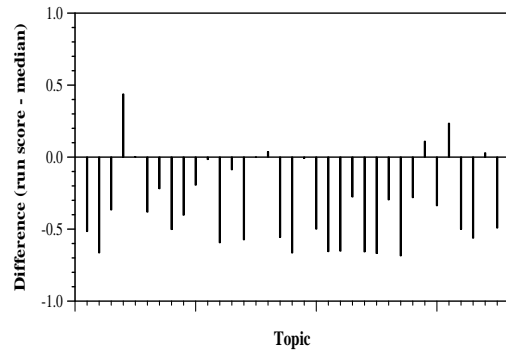


Per-topic difference from median NDCG@10 for all Round 2 runs

Round 2 results — Run factum-dense-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