

Round 1 results — Run dmis-rnd1-run3 submitted from KoreaUniversity_DMIS

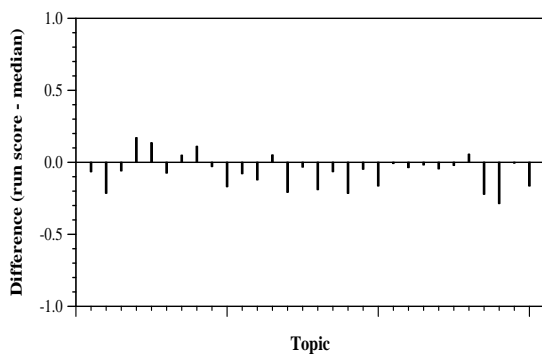
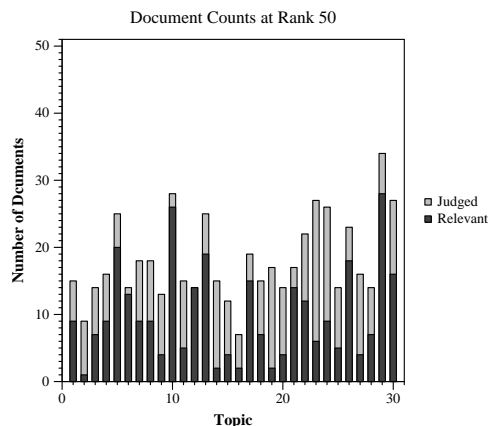
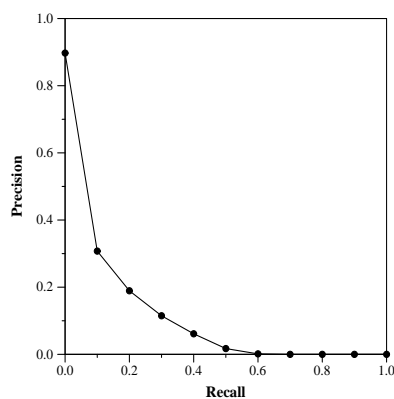
Run Description

We mainly used covidAsk (<https://covidask.korea.ac.kr>), a real-time QA system based on DenSPI [1] for the submission. While the initial purpose of the system was to give answers to natural questions in fine-grained phrases, covidAsk implicitly performs IR as documents that contain correct answer phrases can be regarded as relevant. For this submission, we used only subsets of CORD-19 documents that contain synonyms of 'COVID-19' in their titles or abstracts. This gave us about 3K documents from which we indexed about 800K phrase vectors. As our document representation of each phrase was too simple (BM25), we also combined document scores from Covidex [2]. We found the hyperparameters with our small validation set (100 QA pairs) and chose between 'question' and 'query' in each topic manually as an input to DenSPI trained on SQuAD (Dense-First Search) or SQuAD+NaturalQuestions (Hybrid Search), respectively. Also we manually modified typos and ambiguity in queries. [1] Real-Time Open-Domain Question Answering with Dense-Sparse Phrase Index, Seo et al., 2019 [2] Rapidly Deploying a Neural Search Engine for the COVID-19 Open Research Dataset: Preliminary Thoughts and Lessons Learned, Zhang et al., 2020

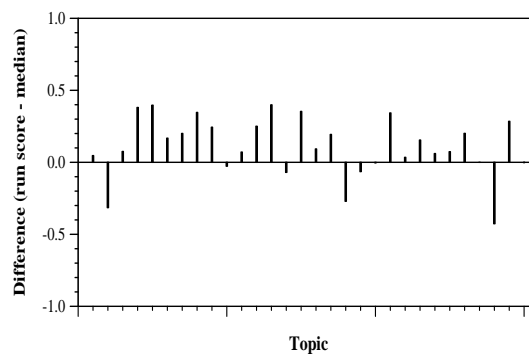
Summary Statistics	
Run ID	dmis-rnd1-run3
Topic type	manual
Contributed to judgment sets?	yes

Overall measures	
Number of topics	30
Total number retrieved	13466
Total relevant	2352
Total relevant retrieved	775
MAP	0.1071
Mean Bpref	0.2601
Mean NDCG@10	0.4649

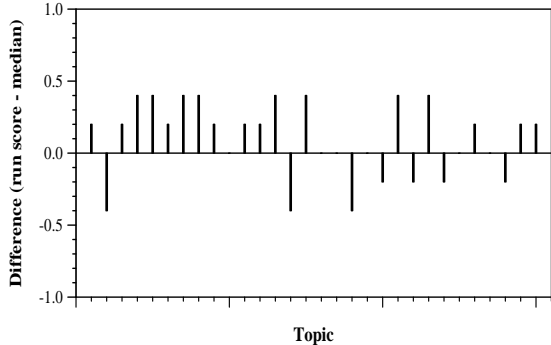
Document Level Averages	
	Precision
At 5 docs	0.5867
At 10 docs	0.4367
At 15 docs	0.3489
At 20 docs	0.3100
At 30 docs	0.2567
R-Precision	
Exact	0.1655



Per-topic difference from median bpref for all Round 1 runs



Per-topic difference from median NDCG@10 for all Round 1 runs



Per-topic difference from median P@5 for all Round 1 runs