Round 2 results — Run cogir-ibm-qQ-PolRnk submitted from CogIR

Run Description

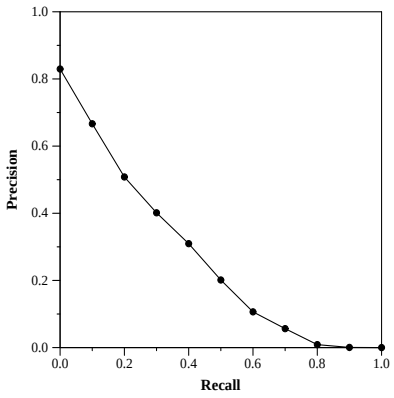
Indexing: Documents were indexed using Lucene version 7.1.0 with EnglishAnalyzer and InQuery stop list. Three fields were kept per document: title, abstract and content. Missing article title or abstract were augmented using the article's content (when available). * Title augmentation: first abstract's text paragraph (from json file). * Asbtact augmentation: + take the longer abstract among metadata and article's JSON + for missing abstract: augmented it with either of the first 500 chars of the article's introduction/summary/conclusions section (in this preference order), + otherwise: if content exists: take the first 500 chars of the content. Articles with no title or abstract were discarded. Total index size: 50309 docs (#Longer abstract taken: 9099, #Abstracts augmented: 844, #Titles augmented: 38). Retrieval: Retrieval was performed in two main phases (baseline rankers and fusion). The first phase was used to derive an intial pool of potential candidates. Here, the topic's query+question parts were used as the searched query. First baseline retrieval phase used three different rankers: 1. Lucene-based retrieval using 4 Lucene similarities: * AxiomaticF1LOG(default) * DFRSimilarity(BasicModelIF,AfterEffectB,NormalizationH3) * LMDirichletSimilarity(mu=200) * BM25Similarity(default) For each similarity the top-1000 documents were retrieved using a multi-field document retrieval: titleˆ0.03,abstractˆ0.05,contentˆ1. Field boosts were tuned using BM25Similarity(default). Various ranked lists were then first fused using CombSUM without score normalization. Following [1] we applied the pseudo-relevance-feedback PoolRank method with CombSUM fusion scores as prior-scores. For pseudo-relevance, we used RM1 [2] with PRF-size=20, Dir-smooth=200, clip size=100 and logistic interplolation (lambda=0.01) tuned using BM25Similarity(default). Following [3] the pool was further reranked using the MaxPsg method (passages were extracted using sliding window of 200 chars with 10% overlap, BM25(K1=0.8,b=0.3) scoring), tuned using BM25Similarity(default). The list of 1000 docs for each query (from the first ranker above), were then re-ranked using the following two fine-tuned BERT models [4]: 2. BERT-Q-a Documents' (title, abstract) pairs were used in the collection as a weak-supervision to fine-tune SciBERT pre-trained model [5] for matching titles to abstracts. Then at run time, given a topic with 1000 docs, document abstracts were scored by matching the topic's question to each abstract using the fine-tuned BERT-Q-a model. 3. BERT-Q-q Similarly, documents' (title, abstract) pairs were used in the collection as a weak-supervision to generate title paraphrases [4]. Those paraphrases were then used to fine-tune another SciBERT model for matching titles to their paraphrases. Then at run time, given a topic with 1000 docs, their titles were scored by matching the topic's question to each title using the fine-tuned BERT-Q-q model. Second fusion phase was applied to combine the three baseline rankers. Following [4], we applied the pseudo-relevance feedback PoolRank method with Weighted CombSUM(weights=[3,2,1],max-min normalization) using again RM1 model (PRF-size=3, Dir-smooth=200, clip size=100 and logistic interplolation (lambda=0.01), tuned using BM25Similarity(default). [1] H. Roitman, Utilizing Pseudo-Relevance Feedback in Fusion-based Retrieval, Proc. of ICTIR'2018 [2] V. Lavrenko, Bruce W. Croft, Relevance-based Language Models, Proc. of SIGIR'2001 [3] H. Roitman, Y. Mass, Utilizing Passages in Fusion-based Document Retrieval, Proc. of ICTIR'2019 [4] Y. Mass, B. Carmeli, H. Roitman, D. Konopnicki, Unsupervised FAQ Retrieval with Question Generation and BERT, to appear in ACL'2020 [5] Iz Beltagy, Kyle Lo, Arman Cohan, SciBERT: A Pretrained Language Model for Scientific Text, CoRR abs/1903.10676 (2019)

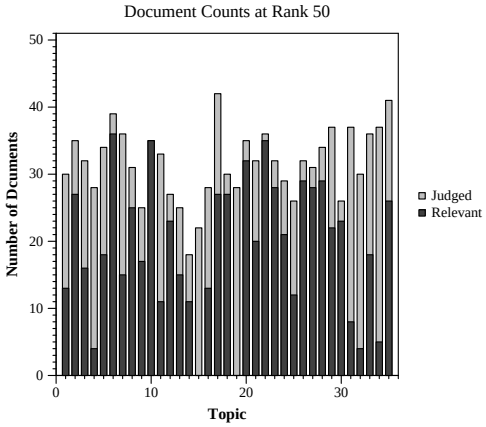
Summary Statistics	
Run ID	cogir-ibm-qQ-PolRnk
Topic type	automatic
Contributed to judgment sets?	yes

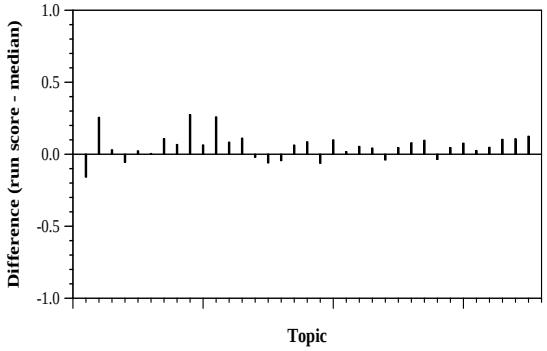
Overall measures	
Number of topics	35
Total number retrieved	35000
Total relevant	3002
Total relevant retrieved	1816
MAP	0.2628
Mean Bpref	0.4250
Mean NDCG@10	0.6104
Mean RBP(p=0.5)	0.6480 +0.0011

Round 2 results — Run cogir-ibm-qQ-PolRnk submitted from CogIR

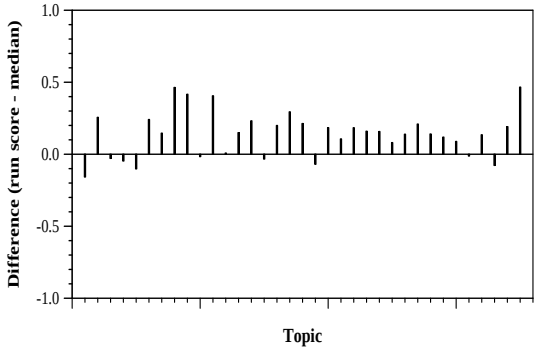
Document Level Averages	
	Precision
At 5 docs	0.7314
At 10 docs	0.6571
At 15 docs	0.6038
At 20 docs	0.5629
At 30 docs	0.4838
R-Precision	
Exact	0.3144



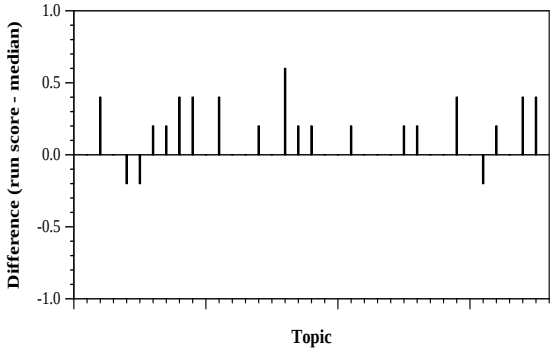




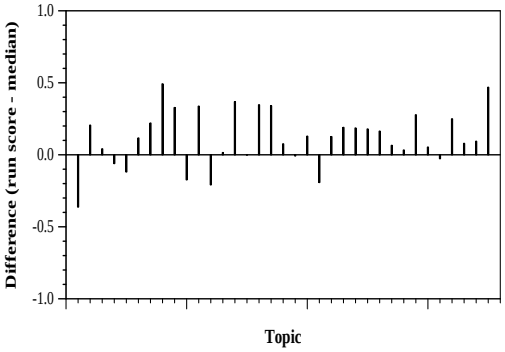
Per-topic difference from median bpref for all Round 2 runs



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



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