exam paper

> Hossein A. Rahmani, Nick Craswell, Emine Yilmaz, Bhaskar Mitra, and Daniel Campos. 2024. Synthetic Test Collections for Retrieval Evaluation. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), July 14–18, 2024, Washington, DC, USA. ACM, New York, NY, USA, 5 pages. arXiv:2405.07767v1

what is a test-collection?

we need test-collections to evaluate systems.

based on the Cranfield-experiments:

- documents
- queries
- relevance-judgements:
    - binary labels (relevant vs. not relevant) / graded labels (score usually between 0;3)
    - sparse / dense
    - implicit feedback / explicit feedback

how do we generate test-collections?

- we need test collections to evaluate systems
- up until now they weren't widely accessible:
    - sampled (ie. ms-marco) → real usage query logs only accessible to search companies
    - handcrafted (ie. trec) → expensive to hire people
- but LLMs can also be used to create synthetic collections:
    - synthetic documents (already explored by other papers)
    - synthetic queries
    - synthetic relevance-judgements

how do we generate synthetic queries?

query = collection of terms to reach a document. should make sense by themselves without any further context.

steps:

- i. sample ~1000 msmarcov2 passages
- ii. filter out low quality passages (around 15%)
    - high query-independent passage-quality-score = passages that don't need any context to be understood
    - score generated by gpt4
- iii. generate queries to which the passage is relevant
    - generate one set with by BeIR-t5 (trained on msmarco)
    - generate one set with zero-shot gpt4
- iv. filter out low quality queries (around 30-40%)
    - done by experts

when compared to real queries:

- differences:
    - contain much fewer relevant documents (score ≥0)
    - are harder to label
    - tend to be longer (from gpt4)
- benchmark: similar performance
    - similar performance of systems both in terms of evaluation results and system ranking
    - synthetic vs. real queries (both with human jugement)
    - τ = 0.8151 (kendall rank corellation)

    - NDCG@10
    - 31 systems

how do we generate synthetic relevance-judgements?

relevance-judgement = mapping of queries to relevant documents, with a label (either binary or graded)

steps: just label documents using gpt4

when compared to real judgements:

- differences:
    - tend to be graded less critically (by gpt4)
- benchmark:
    - using the passage that the query was generated from as the only judgement (= sparse judgement, binary label) is very ineffective: τ = 0.157
    - combining synthetic queries with synthetic judgements yields better results than human judgements: τ = ~0.84-0.85

how useful are synthetic collections?

they can be a very useful extension to existing test-collections:

> It can be seen that evaluation on the fully synthetic test collection results in similar results to human queries with human judgments in terms of system ordering, with a Kendall's τ = 0.86 for NDCG@10

but it's likely that systems evaluated with test-collections that were generated with the same model as they use for retrieval might get favored, leading to biased results. this kind of bias wasn't observed but could be mitigated by combining multiple models.