The Utility of Query Expansion for Semantic Re-ranking Models

An empirical analysis on the performance for ad-hoc retrieval

TUDelft

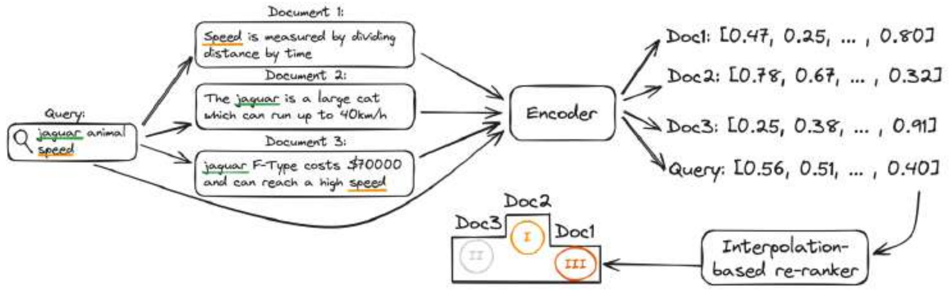
1. Background Information

Ad-hoc retrieval is an Information Retrieval task which identifies and retrieves relevant documents to a user's search query, the most common example being web searches. Initial models, like BM25, proposed exact term matching, based on lexical similarity. These lexical models are fast and efficient, but had limited as they ignore semantic similarities such as synonyms.

To overcome vocabulary mismatch, a solution proposed was **Query Expansion (QE)**, which augments the initial search query with new terms, to capture documents that are relevant. A popular strategy for QE is pseudo-relevance feedback (PRF), which retrieves the most relevant documents, adds relevant terms from these retrieved documents to the initial query and re-runs the retrieval stage with the new query.

Another solution proposed for vocabulary mismatch is the use of Natural Language Models to rank documents based on their semantic similarity of their embeddings into a densely represented vectorial space. Retrieving documents based on their semantic similarity is computationally expensive for large corpus of data, so a hybrid approach is studied during this research: **Semantic Re-ranking Models,** employing a two-stage retrieval pipeline. First,These models use a lexical retrieval to retrieve relevant documents. Then, these documents are re-ranked based on an interpolation of their lexical and semantic similarity.

$$score_{\text{interpolated}} = \alpha \cdot s_{\text{lexical}} + (1-\alpha) \cdot s_{\text{semantic}}$$


Query:
jaguar animal speed
Document 1:
Speed is measured by dividing distance by time
Document 2:
The jaguar is a large cat which can run up to 40km/h
Document 3:
jaguar F-Type costs $70000 and can reach a high speed
Encoder
Doc1: [0.47, 0.25, ... , 0.80]
Doc2: [0.78, 0.67, ... , 0.32]
Doc3: [0.25, 0.38, ... , 0.91]
Query: [0.56, 0.51, ... , 0.40]
Interpolation-based re-ranker
Doc2
Doc3
Doc1
I
II
III

2. Research Question

This research investigates **the utility of Query Expansion for Semantic Re-ranking Models**. More specifically, the following sub-questions will be answered:

1. **Does combining query expansion with semantic re-ranking models lead to better results for ad-hoc retrieval tasks?**
2. **How does the number of retrieved documents in the first stage impact the performance for query expansion combined with a semantic re-ranking model?**
3. **What type of queries benefit from a query expansion stage in a retrieve-and-re-rank pipeline?**

3. Experimental Setup

Models

- **"BM25":** Standard term-frequency model, based on Okapi BM25 [1].
- **"BM25→RM3":** Three-stage lexical retrieval pipeline: PRF QE strategy: BM25 retrieval, RM3 [2] PRF QE stage followed by another BM25 retrieval.
- **"BM25→TCT-ColBERT"** - Two-stage retrieve-and-rerank pipeline: First, BM25 is used to retrieve the most relevant documents. Then, the documents are re-ranked based on the interpolated score. As the encoder, **TCT-ColBERT** [3] model is used for both documents and queries. For a faster retrieval, the embeddings of the entire corpus are pre-computed and stored using Fast-Forward Index [4], from which the documents' embeddings are retrieved at runtime.
- **"BM25→RM3→TCT-ColBERT"** - a combination of the previous two models, replacing the BM25 retrieval stage of TCT-ColBERT with a RM3 QE strategy.

Datasets & Tasks

For a representative result, the tasks and datasets used are standard benchmarks for ad-hoc retrieval. These datasets are extracted from **BEIR** (**BE**nchmarking **I**nformation **R**etrieval) and **TREC** (**T**ext **RE**trieval **C**onference) Deep Learning Track. For this research, 20 different datasets were used, varying in both subject (eg. general topics Q&A, financial, medical) and task (eg. argument retrieval, duplicate question retrieval).

Evaluation Metrics

Three different metrics were used to evaluate the performance of the models:

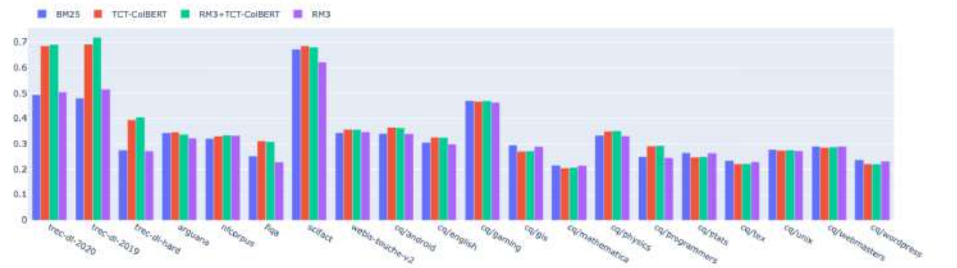
- **"RR"** - (Mean) Reciprocal Rank: precision-focused measurement, which associates a score based on the position of the first relevant document
- **"MAP"** - Mean Average Precision: measures the average precision of the retrieved documents in the set of ranked results.
- **"nDCG"** - Normalised Discounted Cumulative Gain: scores (document, query) pairs retrieved against an ideal ranking of relevant documents.

5. Limitations & Future Work

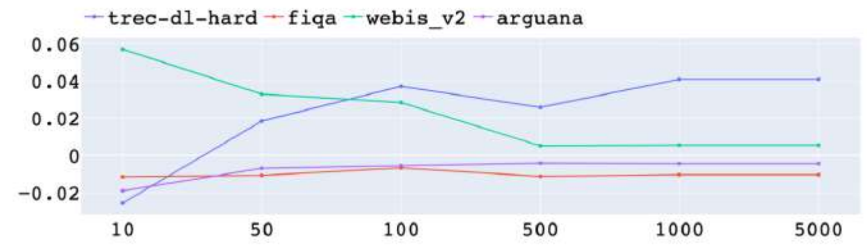
The results obtained during this research are constrained by two factors: time and resources. Given the **limited time** of the project and the **hardware capabilities** available, this research only focused on one QE strategy and one semantic re-ranking model, tested against a subset of the benchmarks available.
For a more comprehensive and detailed conclusion, it is advised to expand on both the **models** used, analysing the performance on newer and more performant models, as well as **datasets**, given that more complex ones are available, such as *'msmarco-passage-v2'*, which orders of magnitude larger than the ones used in this research.

4. Results

RQ: *Results after running all models against the 20 datasets, using the "nDCG @ 10" metric*



RQ2: *X-axis displays the number of retrieved documents, Y-axis displays the difference ΔRR, where ΔRR is the difference in "RR @ 10" score between RM3+TCT-ColBERT and TCT-ColBERT*



RQ3: *Example of a query where RM3 lead to a ΔRR of 0.875. The difference in lexical score ranked the most relevant document the 1st for RM3+TCT-ColBERT and 8th for TCT-ColBERT*

fiqa: "for insurance why should you refuse 4 000 year for only 10 years and prefer 500 year indefinitely" (Query ID: 5155)

indefinite(0.100) 10(0.096)
salari(0.048) year(0.050)
current(0.022)
hous(0.036) 000(0.090)
onli(0.050) average(0.060)
500(0.091)
scenario(0.027)
prefer(0.050) refuse(0.05)
insur(0.075) 4(0.050)
	"The breakeven amount isn't at 8 years. You calculated how many years of paying $500 it would take to break even with one year of paying $4000. 8 x 10 years = 80 years. So by paying $500/year it will take you 80 years to have spent the same amount ($40000 total) as you did in 10 years. At this point it may seem obvious what the better choice is. Consider where you'll be after 10 years: In scenario 1 you've spent $5000 ($500*10) and have to continue spending $500/year indefinitely. In scenario 2 you've spent $40000 ($4000*10) and don't have to pay any more, but you currently have $35000 ($40000 - $5000) less than you did in scenario 1. [...]" (Document no: 462892)

6. References

1. S. E. Robertson, S. Walker, S. Jones, M. M. Hancock- Beaulieu, and M. Gatford, "Okapi at trec-3," NIST Spe- cial Publication SP, vol. 109, p. 109, 1995.
2. N. Abdul-Jaleel, J. Allan, W. B. Croft, F. Diaz, L. Larkey, X. Li, et al. Umass at trec 2004: Novelty and hard. Computer Science Department Faculty Publication Series, 189, 2004.
3. S.-C. Lin, J.-H. Yang, and J. Lin, "Distilling dense rep- resentations for ranking using tightly-coupled teachers," 2020.
4. J. Leonhardt, K. Rudra, M. Khosla, A. Anand, and A. Anand, "Efficient neural ranking using forward indexes," in Proceedings of the ACM Web Conference 2022, ser. WWW '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 266–276. [Online]. Available: https://doi.org/10.1145/ 3485447.3511955

Victor-Filip Ghita - v.f.ghita@student.tudelft.nl | Supervisor: Jurek Leonhardt (l.j.leonhardt@tudelft.nl) | Responsible Professor: Avishek Anand (avishek.anand@tudelft.nl)