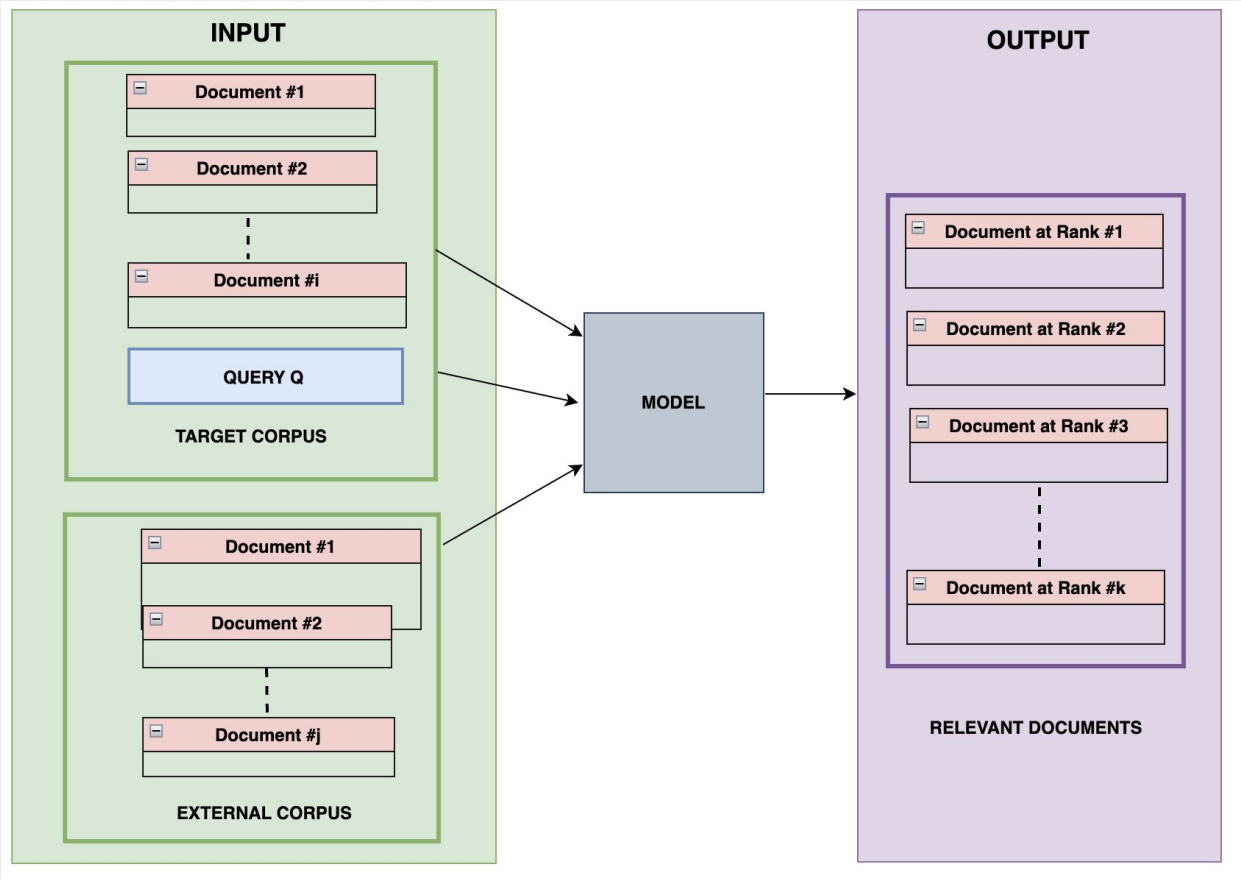
Effect of increasing the number of Feedback Documents for Pseudo-Relevance Feedback on Dense External Expansion for Zero-shot Retrieval

Group 3
Indrajeet Devale, Mona Anil Udasi, Ramprasad Kokkula

Paper: Improving zero-shot retrieval using dense external expansion

Authors: Xiao Wang, Craig Macdonald, Iadh Ounis

Retrieval Task



Zero-shot Retrieval

- Zero-shot retrieval is when a model retrieves relevant items from a dataset without having been explicitly trained on specific examples of the query type or category.

- This approach is called "zero-shot" because the model has to generalize from training data involving different, but somehow related, tasks or categories to handle new, unseen queries.

Target Corpus

BEIR (Benchmarking IR) NFcorpus dataset - This dataset has queries and documents related to nutritional information, making it particularly useful for testing retrieval systems in the context of health and diet-related searches.

Query: "What are the health benefits of olive oil?"

Passage Excerpt 1: "Olive oil is rich in monounsaturated fats, which are considered healthy fats. It has been associated with a lower risk of heart disease and a reduction in cholesterol levels."

Passage Excerpt 2: "The oil contains antioxidants that are believed to have anti-inflammatory properties. These antioxidants can help reduce oxidative stress and may lower the risk of chronic diseases."

External Corpus

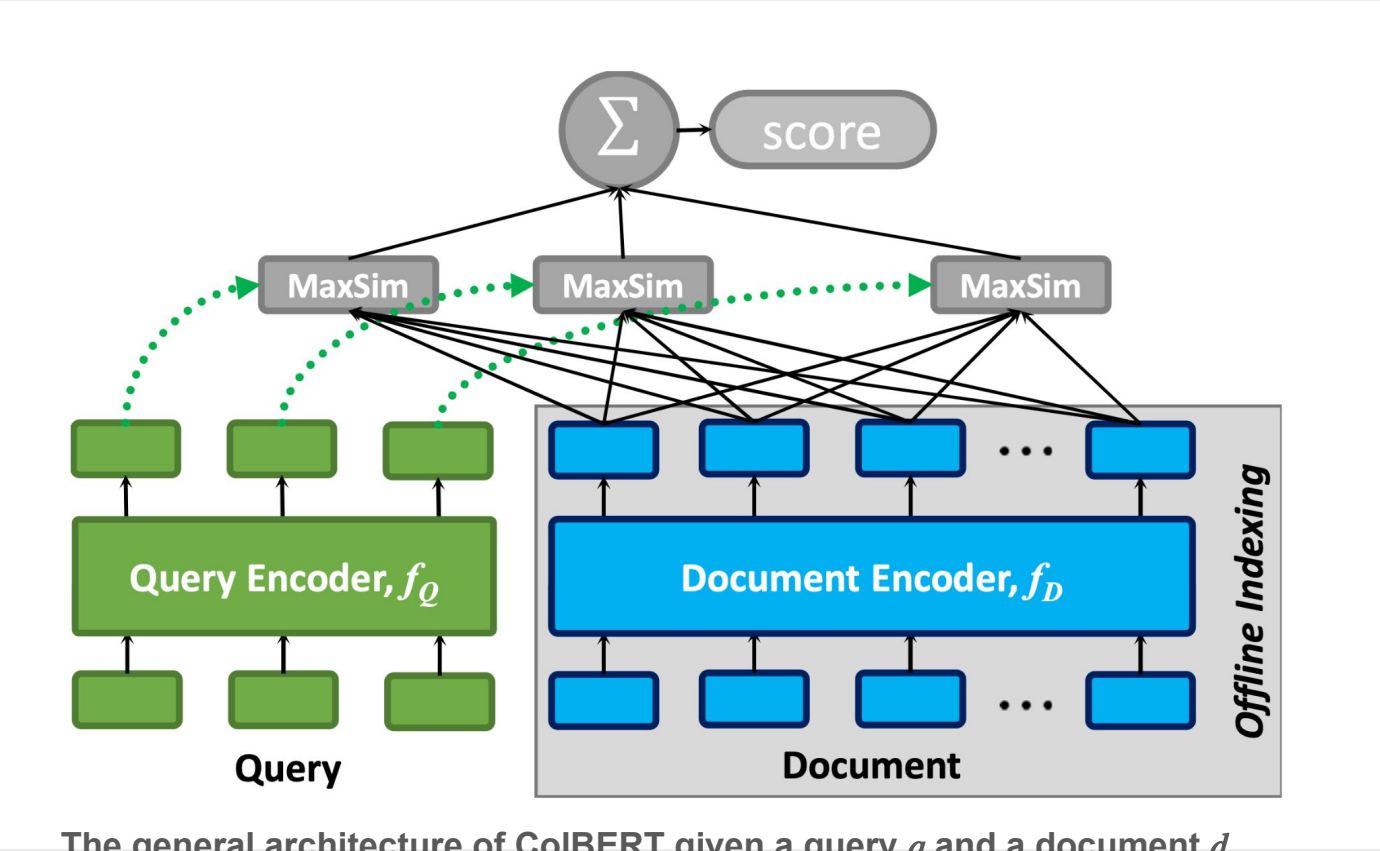
MS MARCO dataset - A large-scale dataset created by Microsoft widely used in the natural language processing community due to its realistic queries and detailed answers with 8.8M documents derived from user questions submitted to the Bing search engine.

Query: "What is a corporation?"

Passage Excerpt 1: "A corporation is a company or group of people authorized to act as a single entity (legally a person) and recognized as such in law."

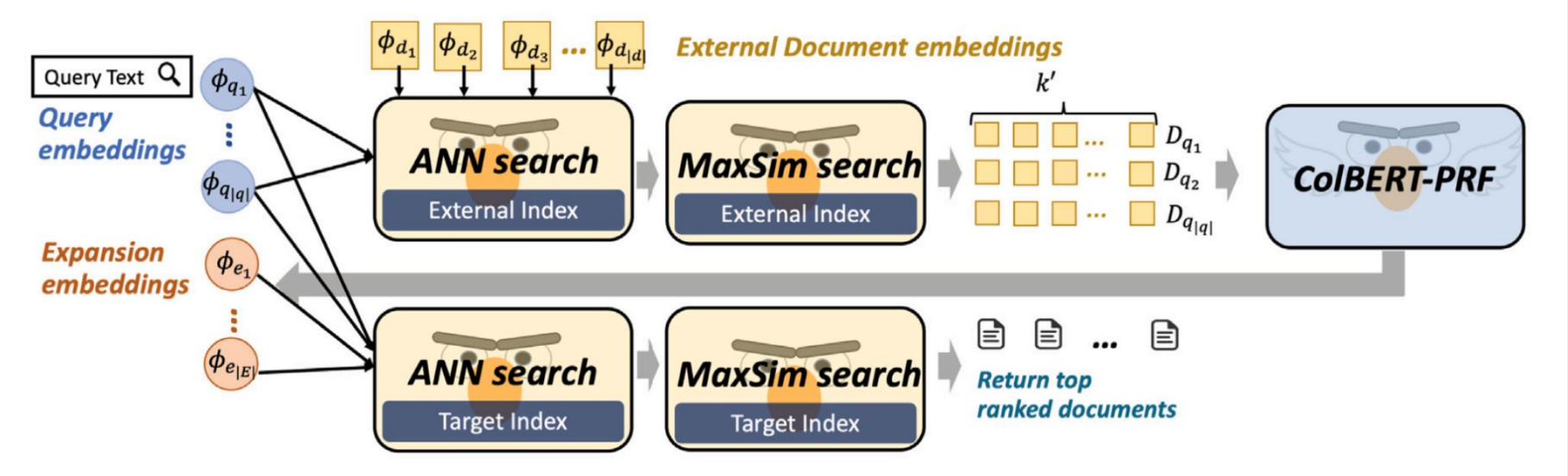
Passage Excerpt 2: "Early incorporated entities were established by charter (i.e., by an ad hoc act granted by a monarch or passed by a parliament or legislature)."

Model



The general architecture of ColBERT given a query q and a document d

Query Expansion Model



Logical Flow of ColBERT-PRF for external expansion

Query Expansion Steps

1. Index the target corpus and external corpus
2. Get topics from the target corpus
3. Perform end-to-end dense retrieval on target corpus using ColBERT to retrieve top-k documents
4. Perform Pseudo-Relevance Feedback using ColBERTPRF and target corpus to get dense query embeddings
5. Perform end-to-end dense retrieval on external corpus using ColBERT to retrieve top-k documents
6. Perform Pseudo-Relevance Feedback using ColBERTPRF and external corpus to get dense query embeddings
7. Concatenate the above query embeddings in a dataframe and perform end-to-end dense retrieval to retrieve top-k documents on target corpus
8.

Results

Original Query - breast cancer and diet

Expanded query - 'ap', 'diagnosis', 'calculated', 'vegetables', 'breast', '95', '6', 'cancer', 'intake', 'dietary'.

Experimental Setup

- Pyterrier
- PyTerrier ColBERT plugin for indexing and PRF
- Pandas
- Target Corpus - BEIR NFCorpus
- External Corpus - MS MARCO

Research Question

- What is the impact on effectiveness if the number of feedback documents considered in the external corpus are increased while performing pseudo-relevance feedback?

Independent Variables

- Number of feedback documents considered in the external corpus
- Beta: Variable to control the influence of external embeddings on original query

Control Variables

- Set of queries from the target corpus
- Target and External Corpora
- Number of clusters considered while performing PRF

Dependent Variables

- nDCG_cut_10: normalized Discounted Cumulative Gain at 10
- MAP: Mean Average Precision

Hypothesis

- The effectiveness should increase when the number of feedback documents considered in the external corpus are increased while performing pseudo-relevance feedback

Table 1: External expansion for Multiple Representation Dense Retrieval

Beta	Feedback Documents	NDCG@10	NDCG@20	NDCG@1000	MAP
0	3	0.31166	0.28850	0.35870	0.15233
0	5	0.31166	0.28850	0.35859	0.15231
0	10	0.31166	0.28850	0.35858	0.15236
0	15	0.31166	0.28850	0.35846	0.15233
0	20	0.31166	0.28850	0.35867	0.15234
0	25	0.31166	0.28850	0.35859	0.15234
0	30	0.31166	0.28850	0.35875	0.15236
0.5	3	0.31011	0.28703	0.35948	0.15251
0.5	5	0.31056	0.28555	0.35701	0.15013
0.5	10	0.31120	0.28672	0.35874	0.15183
0.5	15	0.31212	0.28897	0.35901	0.15222
0.5	20	0.30964	0.28723	0.35763	0.15134
0.5	25	0.31321	0.28842	0.35728	0.15148
0.5	30	0.30973	0.28784	0.35712	0.15173
1	3	0.29191	0.27105	0.34701	0.14094
1	5	0.30024	0.27744	0.35132	0.14601
1	10	0.30416	0.28204	0.35490	0.14904
1	15	0.30343	0.28340	0.35420	0.14978
1	20	0.30532	0.28361	0.35441	0.14903
1	25	0.30825	0.28478	0.35452	0.14878
1	30	0.30546	0.28399	0.35352	0.14952

Comparison Against Baselines

Table 2: NDCG and MAP metrics for different search models

Name	NDCG_Cut_10	NDCG_Cut_100	NDCG_Cut_1000	MAP
BM25	0.322219	0.272919	0.300229	0.148882
BM25 + RM3 Query Expansion	0.334130	0.303094	0.372693	0.167958
DPH	0.313269	0.266587	0.295735	0.145268
DPH + Bo1 Query Expansion	0.331988	0.302251	0.369949	0.170919
Zero-shot Retrieval ($\beta = 1$, fbdocs $= 3$)	0.29191	0.27105	0.34701	0.14094
Zero-shot Retrieval ($\beta = 1$, fbdocs $= 30$)	0.30546	0.28399	0.35352	0.14952

Limitations

1. Our external corpus is not large enough to capture all the relevant token embeddings from the external corpus pertaining to the original query.
2. Additional noisy embeddings can sway the expected topic of the retrieved documents.
3. Having an overpowering Beta value, which affects the weight of the expansion term embeddings can lead to a completely off-topic different set of retrieved documents.

Thank You!
Questions?