A Distribution-Aware Approach to Dense Retrieval

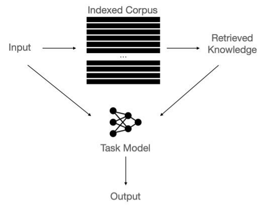
Jason Lin, Justin Young, Simran Arora
{jj0, justiny, simarora}@stanford.edu

Stanford | ENGINEERING
Computer Science

Problem

Information Retrieval (IR) is an important step for open-domain applications such as language modeling, question-answering, fact-checking, and personal assistants. These systems are often evaluated on one distribution, while many real world applications of retrieval require retrieving simultaneously from multiple distributions. As most existing textual benchmarks for QA involve retrieving from one distribution, we first consider how to set up and characterize a multi-distribution retrieval setting, and next strategies for retrieving from both distributions.

Background



Methods

To decouple the confounding factor that a question can be answered by passages from multiple domains combined, we create synthetic subpopulations within MSMARCO and finetune with distribution-aware training strategies.

Unsupervised Synthetic Domain Split: We use UMAP for non-linear dimensionality reduction before clustering corpus documents with K-means. To set the number of clusters, we use the Gap statistic approach which intuitively captures how tight points are around a cluster. This forms the basis for our ID (A) and OOD (B) split.

Training Strategies: Prior work does not fine-tune on one distribution (A) and evaluate relevance on a different distribution (B). We perform an exhaustive set of tests where we fine-tune on all combinations of A and B and examine the subsequent relevance scores on queries from the A or B test sets. We also examine the effects of fine-tuning with BM25 hard negatives evaluated on said distributions.

Data

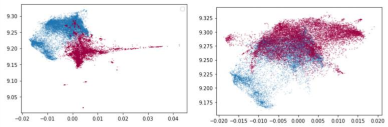
We pull data from MS MARCO, a passage-ranking dataset of Bing user queries and relevance passages from multiple web sources.

- 8,841,823 million passages.
- 532,761 (query, passage) pairs in *train* set
- 6,980 *test* queries.

For computational feasibility, we pull 50k (query, passage) pairs from the train set, all of the 6,980 test pairs, and an additional 150k random passages.

Clusters

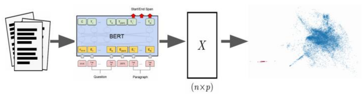
Blue is our "home" cluster. **Left**: cluster furthest from home. **Right**: cluster closest to home. Plotting the 30-dimensional UMAP embeddings along two random axes. Number of clusters 23 is selected by a GAP statistic: $Gap_n(k) = E_n^*\{\log(W_k)\} - \log(W_k)$



Cluster Quality

Semantically similar topics bunch together! Both **by passage and by queries.**

```
does schizophrenia cause hallucinations
vasospasms caused by what
cad heart related
what are aneurysm
what are signs of anxiety in your chest
what are the complications of varicose vein
can a pinched nerve cause tooth pain
what are the early signs of colon cancer?
what are the most common causes of paralysis
what are the signs of allergies in the winter time
what are the signs of kidney failure in dogs with dm?
what can be reason of excess urination
what cause dizziness mayo clinic
```

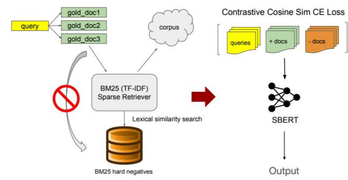


Basic pipeline: encode, UMAP, cluster

Experiments

We start with a pre-trained distilBERT bi-encoder model on the full MS MARCO data, eliminating the bias toward any particular subset of MS MARCO. We compare the following fine-tuning baselines: (1) fine-tuning on questions corresponding to the in-distribution training questions and evaluating on **both ID and OOD** test questions, (2) fine-tuning on a question set of the same size as (1), but randomly splitting the training questions, (3) no fine-tuning, and (4) fine-tuning with BM25-mined hard negatives. Scores reported are NDCG@10. Finetuning (1), (2) uses in-batch negatives.

Fine-tuning Regime	ID Test Queries	OOD Test Queries
ID training queries	0.6417	0.6619
Training queries, random split	0.6292	0.6843
No tuning	0.7710	0.8334
BM25-mined hard negatives	**0.7767**	**0.8371**



Analysis

1. BM25 achieves high relevance scores and is a strong baseline for sparse retrieval. **Outperformance is robust across different distribution splits**. Even though our model is pretrained on the full MSMARCO (8.8M), by fine-tuning a dense retriever on a small cluster (30K) with hard negatives indexed from a 200K corpus subset, we can further improve ID and OOD retrieval.
2. Vanilla fine-tuning on subsets actually degraded performance vs. no fine-tuning, but was better than fine-tuning on random query data for in-distribution splits
3. We observe **a larger percent of passages retrieved for OOD questions are actually** OOD passages, compared to the percent of ID passages retrieved for ID questions. This does not support our hypothesis that a retriever trained on a biased sub-distribution might favor ID passages for OOD questions.

Conclusion

Dense retrievers are highly sensitive to the training strategy and data selection. We proposed a novel method to construct synthetic multi-distribution retrieval settings, showed that vanilla fine-tuning can degrade performance, and that BM25 fine-tuning is consistently helpful for generalization.

For existing retriever training datasets, corpora are often orders of magnitude larger than the training datasets used to train the question and passage BERT encoders. For example, on the MS MARCO benchmark, the number of documents in the corpus is 18x larger than the number of training pairs, so several documents are not incorporated during training process. Our setting, conditioned on further investigation, could implications for how to design question answering benchmarks with good coverage over question and passage types. **Overall, we hope this work encourages further attention towards retrieval strategies that account for the sub-distributions in the background corpus.**

Limitations

Understanding Degradation: As a priority, we need to pinpoint why degradation occurs for the ID test queries. We speculate it has to deal with overtraining on a very specific subset of the train queries.

Characterizing the clusters: The metrics we use to characterize the clusters currently are limited to centroid norm and pairwise distance between embedded passages. Perhaps other metrics exist that can inform us about how relevance scores change with observable cluster characteristics.

Future Work

1. **Synthetic Split Extensions**
   a. Alternative creation methods
   b. Increasing the data size
   c. Robustness checks on results based on "home" cluster chosen
2. **Optimizing for ID and OOD performance without degradation**
   a. Distributionally Robust Optimization (DRO)
   b. LoRA: Fine-tune longer with a low-rank adaptation of the retriever's underlying BERT-based language model
   c. LP-FT: Linear Probing then Fine-Tuning to mitigate ID-OOD tradeoffs

References

[1] Nandan Thakur, Nils Reimers, Andreas Ruckle, Abhishek Srivastav, and Iryna Gurevych. Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. Thirty- fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS), 2021.

[2] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Towards unsupervised dense information retrieval with contrastive learning, 2021.

[3] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert- networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (ACL), 11 2019.