Multi Distribution Dense Information Retrieval

Stanford CS224N Custom Project

Soumya Chatterjee
Department of Computer Science
Stanford University
soumyac@stanford.edu

Abstract

In this project, we propose the novel problem of multi-distribution information retrieval where given a query we need to retrieve passages from different corpora, each drawn from a different distribution. Some of these distributions might not be known at train time. This is a very natural setting that arises for queries such as "Is Mbappé younger than me?" which require retrieval from public and private data sources – personal information, medical data or data private to organizations. Due to the lack of existing benchmarks on this novel setting, we design benchmarks for this task by adapting existing multi-hop question answering and entity matching datasets. Specifically, we create one question answering and two entity matching based datasets for evaluating multi-distribution retrieval. We propose simple methods for this task which allocate the fixed retrieval budget (top-k passages) strategically across domains to prevent the known domains from consuming most of the budget. We show that our methods lead to 8+ point improvements in Recall@100 over three datasets and that improvements are consistent when fine-tuning different base models.[1]

1 Key Information to include

- Mentor: Eric Frankel
- External mentors/collaborators: Simran Arora (simranarora@stanford.edu), Omar Khattab (okhattab@stanford.edu)
- Sharing project: Related to my ongoing research project. Details in proposal.

2 Introduction

Open-domain Question Answering (QA) is a ubiquitous NLP problem with applications in search, personal assistants and customer service, among others. It involves building systems the can understand questions asked in natural language and provide accurate and relevant answers. Open-domain QA is challenging due to the diversity of the questions that can be asked and the vastness of the information that needs to be searched. One class of approaches to open-domain QA is retrieval-based systems which work by selecting the most relevant passages from a corpus given a query which are used to find the answer. The performance of the retrieval system in critical in these approaches.

In open-domain QA, some queries might be such that answering them requires information from multiple data sources like the recent news and Wikipedia articles. For such questions, retrieval systems that are capable of retrieving from different sources is needed. Further, training might not be feasible on some sources due to reasons including privacy or restricted access. In these cases, the

[1] Other than privacy reasons, the multi-distribution setting could occur when retrieving from data that has not been created at the time of training. For example, for questions which need passages from Wikipedia and news articles created in the future (about events that have still not happened at training time).

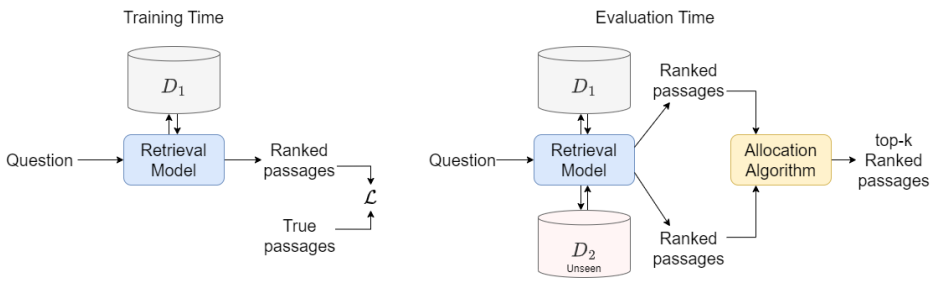

Figure 1: Overview of the multi-distribution retrieval task. During training some distributions $\mathcal{D}_1$ are available and during evaluation, queries which require retrieval from the seen distributions $\mathcal{D}_1$ and unseen distributions $\mathcal{D}_2$ are asked. Different allocation algorithms are used to retrieve a fixed number k of passages by combining the retrieved passages from $\mathcal{D}_1$ and $\mathcal{D}_2$. This is needed so that passages from $\mathcal{D}_1$, which the model was trained on, do not consume most of the budget k.

retrieval system trained on some domains should be able to perform well on other unseen domains too. We formalize this requirement as multi-distribution information retrieval where some of the distributions are unseen during training. This is a challenging task since the retriever needs to be able to generalize to new domains during testing. The problem is different from that of domain adaptation where target domain examples are available. The wide applicability of this setting in domains such as healthcare, finance, enterprise search, etc., make it an interesting and challenging task to study.

To enable further research on this topic, we create three benchmarks, namely Walmart-Amazon, Amazon-Google and a modified version of CONCURRENTQA based on existing entity matching (Das et al., 2017) and multi-hop question answering (Arora et al., 2022) datasets. These are challenging datasets with oracular approaches achieving up to 10 point higher recall than simple baselines. We propose simple approaches for the multi-distribution retrieval task which allocate the fixed retrieval budget across distributions leading to about +3.85 recall over baselines, on average.

Contributions

- We propose the novel task of fixed-budget multi-distribution retrieval where certain distributions are unseen at training time. This problem setting arises naturally due to privacy and other reasons.
- We create three benchmarks for this task based on entity-matching and question answering datasets.
- We experiment with simple methods for multi-distribution retrieval that strategically allocate the fixed retrieval budget across distributions and report performance on the benchmarks created.
- We perform thorough analysis and ablation studies to gain insight into the benchmarks and allocation strategies. We investigate the effect of size of retrieval budget, size of training set, and choice of base pretrained models.

3 Related Work

Open-Domain Question Answering: Question Answering (QA) is a widely studied NLP task that can be categorized into extractive or generative QA depending on whether the answer is a span of given passages or it needs to be synthesized. It can also be categorized into reading comprehension where the question are asked about a given passage and open-domain QA where the answer sources are not restricted to a passage (it can be the entirely of Wikipedia, the internet or simply commonsense). Some works also formulate QA as multi-choice questions. Several benchmarks have been proposed for QA ranging from SQuAD (Rajpurkar et al., 2016), Natural Questions (Kwiatkowski et al., 2019) to HotpotQA (Yang et al., 2018) and many others (Bajaj et al., 2016; Joshi et al., 2017; Dua et al., 2019; Zhang and Choi, 2021; Pang et al., 2022) to evaluate various aspects of QA systems as have been several systems including DPR (Karpukhin et al., 2020) and BiDAF (Seo et al., 2017). Several general-purpose large language models including GPT (Radford et al., 2019), T5 (Raffel et al., 2020), etc are also capable of question answering either by in-context learning (Brown et al., 2020), instruction fine-tuning (Raffel et al., 2020; Wei et al., 2022) or task-specific fine-tuning. In this project, we deal with the retrieval step of open domain question answering where documents which might contain the answer to a question are selected from a large corpus. We fine-tune pretrained

Transformer encoder models (Vaswani et al., 2017) like BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019b) using data from the known distributions.

Retrieval-Based Systems: Several NLP applications including open-domain question answering (QA) (Voorhees et al., 1999) and personal assistants (Nehring et al., 2021) require the ability to handle a wide range of topics. This capability is usually added in one of two ways – implicit-memory methods which make model parameters 'remember' information, and retrieval-based methods which learn to fetch relevant documents from a corpus (like Wikipedia) or a knowledge graph. Our project focuses on retrieval-based methods. Retrieval-based systems typically consist of a retriever and a reader (Chen et al., 2017). The retriever performs maximum inner product search over embeddings of the question and all documents. The reader is a separate model that takes in the question and retrieved documents to generate the answer (typically autoregressively). Related to our task, Arora et al. (2022) created a dataset CONCURRENTQA for retrieval from private and public corpora while satisfying privacy constraints. A version of this dataset modified to suit the multi-distribution retrieval task is one of the benchmarks we propose in the project.

Domain Adaptation: Our problem setting bears resemblance to those of domain adaptation (Ben-David et al., 2010) and out-of-domain generalization. In domain adaptation, a model trained on one domain or distribution is adapted to work well on another with the goal of transferring knowledge from a source domain with sufficient labelled data to an unlabelled target domain. This is challenging since neural networks are sensitive to distribution shifts. Several methods like learning domain independent features (Ganin et al., 2016) or adapting based on unlabelled target domain examples (Liu et al., 2019a) have been proposed. However, in our multi-distribution retrieval setting, no examples from the unseen distributions are available and unlike domain adaptation, we want our models to perform well on *all* distributions instead of just the target.

4 Approach

In this section, we define the Multi Distribution Retrieval Task, give an overview of dense retrieval approaches and describe our method for multi-distribution retrieval.

4.1 Multi Distribution Retrieval Task

In this project, we consider the problem of multi-distribution information retrieval where the retrieval corpora comes from different distributions, only a subset of which is available during model training. For example, we might have two data distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ where $\mathcal{D}_1$ is known during training while $\mathcal{D}_2$ is not (Figure 1). Possible reasons for $\mathcal{D}_2$ not being available include it being a private dataset on which we cannot train or it being generated in the future (after the model has been trained).

Specifically for our project, the two distributions are two sources of text passages with different characteristics. For example, one distribution could be Wikipedia passages and the other could be email snippets. The passages from these two sources can be expected to form two different distributions since the encyclopediac nature of Wikipedia passages would be different from the conversational nature of emails. In this example, it could be that example emails might not be available during training due to privacy reasons.

Now, to formalize the task for two distributions and textual passage retrieval, let us say that we have sets of text passages $D_1 = \{d_1^1, d_2^1, \ldots d_m^1\}$ and $D_2 = \{d_1^2, d_2^2, \ldots d_n^2\}$ with different characteristics (e.g. Wikipedia passages and email snippets) drawn from $\mathcal{D}_1$ and $\mathcal{D}_2$ respectively. Here only D_1 is available during training while D_2 is not. Given a query q, the multi-distribution retrieval task requires retrieving two passages d_i^1 and d_j^2 from the two corpora D_1 and D_2 respectively which are most relevant to the given query q. Examples of queries and relevant passages are shown in Table 1.

4.2 Dense Retrieval Methods

For dense retrieval from a corpus $\mathcal{C}$ based on a query q, prior work (e.g. Chen et al., 2017, Karpukhin et al., 2020) use encoders E_Q and E_P to get the embeddings of the query ($\mathbf{e}_q$) and all the passages $\{\mathbf{e}_c \mid c \in \mathcal{C}\}$ respectively. The similarity between the query embedding and each of the passage embeddings in computed and the passages with the highest similarity is returned as an answer to the

Query	Do ShopBack and Cognitive Arts both deal with internet based services?
Passage1 (Wiki)	ShopBack is a Singaporean-headed e-commerce startup that utilises the cashback reward program. It allows online shoppers to take a portion of their cash back when they buy products through ...
Passage2 (Email)	Cognitive Arts, a developer of Internet-based products and services for educational and corporate training uses, said it appointed Russell C. White as chief executive officer ...
Query	How do the Walmart and Amazon listings of Acer Iconia Tablet Bluetooth Keyboard differ?
Passage1	Acer Iconia Tab Bluetooth Keyboard Bluetooth model 2.0 Removable AAA Battery is included 32.8 operating distance LED power pairing battery indicator Thin stylish design Convenient ...
Passage2	The Bluetooth Keyboard for the Acer Iconia Tab is the perfect accessory for increased productivity. Wirelessly connect to your Tab for seamless typing and navigation. This slim keyboard is the perfect travel companion for when you take your Tab on the road. It conveniently fits in ...
Query	Are Eros International and MicroEmissive Displays in the same type of industry?
Passage1 (Wiki)	Eros International PLC is a leading global company in the Indian film entertainment industry, the Isle of Man. Through its production and distribution subsidiary, Eros International, it ...
Passage2 (Email)	MicroEmissive Displays, which develops microdisplays for embedding into portable electronics products, said it raised GBP 1 million ($2.1 million) in its first round of funding ...
Query	Which of American Fur Company or Hyperchip Inc. was founded first?
Passage1 (Wiki)	The American Fur Company (AFC) was founded in 1808, by John Jacob Astor, a German ...
Passage2 (Email)	Richard Norman, president and CTO of Montreal-based Hyperchip Inc ... since co-founding the company back in 1997, he has averaged about 100 hours a week ...

Table 1: Examples of queries and relevant passages from our datasets.

query q. The similarity here can be dot product or cosine similarity. Further, typically, top-k passages (for some k) are returned instead of one. The query and passage encoders E_Q and E_P usually have the same architecture but may or may not have the same parameters (Karpukhin et al., 2020). It is common to use pretrained Transformer encoders (Devlin et al., 2019; Liu et al., 2019b) and fine-tune them for retrieval. The retrieved passages are typically fed to a reader model to generate the answer (Chen et al., 2017) but in the project, we focus on only the retrieval part of the process.

In our work, we fine-tune RoBERTa encoders (Liu et al., 2019b) with parameters shared between E_Q and E_P. In particular, our retrievers are based on those of Xiong et al. (2021) but modified to our single-hop multi-distribution retrieval task. Specifically, we modified the data loaders, loss functions, training loop and implemented different allocation strategies. While Xiong et al. (2021) is designed for multi-hop QA, we found it to be a suitable starting point since Arora et al. (2022), whose dataset we adapt to create a benchmark, was based off the same system.

Training: As discussed in the task formulation, during training, we have access to data from distribution $\mathcal{D}_1$ only in the form of the corpus D_1. For this distribution, we also have access to a dataset $Q_1 = \{(q_1, p_{11}, p_{12}, \ldots), (q_2, p_{21}, p_{22}, \ldots), \ldots\}$ of queries q_i and corresponding relevant passages $\{p_{i1}, p_{i2}, \ldots\}$. Such datasets are commonly used in question answering and are readily available. We use this dataset for fine-tuning our models. For each example in the dataset, we have a query q_i. We sample one positive passages p_{ij} from the positive passages corresponding to q_i and also randomly sample another passage p'_{ij} from the corpus D_1 to act as a negative passage. Using these, the loss on a single example is given by:

$$\mathcal{L}_{ij} = -\log \frac{e^{E_Q(q_i)^\top E_P(p_{ij})}}{e^{E_Q(q_i)^\top E_P(p_{ij})} + e^{E_Q(q_i)^\top E_P(p'_{ij})}}$$

The average of $\mathcal{L}_{ij}$ over a batch of B samples is minimized in one training iteration. Other more sophisticated choices of negative examples like sampling hard-negatives using BM25 can also be employed instead for randomly sampling p'_{ij}.

Evaluation: During evaluation, we are given queries q which need to retrieve two passages p_1 and p_2 from the two corpora D_1 and D_2 representing the two distributions. Also, recall that we return the top-k documents based on the similarity of their embeddings to the query embedding. Now, since the encoders were trained on examples from $\mathcal{D}_1$, we expect it to be proficient in retrieving examples from D_1. However, the encoders, which had not seen passages from $\mathcal{D}_2$ during training, will not be so good in retrieving from D_2. That is, the passages from D_1 would get assigned higher similarity scores than those from D_2.

A simple approach to multi-distribution retrieval would be to combine the two corpora D_1 and D_2 to a single corpus D_{merged} and retrieve from it. However, based on the above observations, we can see that this naïve approach will not work since the passages from D_1 will have higher scores are use

up most of the budget of k passages that we had. A reasonable method here would try to balance between the two corpora and allocate portions of the retrieval budget to both D_1 and D_2. This is because the model would be better at retrieval from D_1 and have less uncertainty in its predictions than that on D_2. To address this issue, we propose various allocation strategies below.

Allocation Strategies: Given a retrieval budget of k passages, several allocation strategies, some novel, described below can be used.

- Naïve merging: Merge the two corpora D_1 and D_2 into a single corpus D_{merged} and retrieve the top-k passages from it. This is the simplest approach and is equivalent to (incorrectly) assuming that the passages D_2 from the unseen distribution are also drawn from the same distribution as D_1.
- Fixed-fractional allocation: Retrieve k_1 passages from D_1 and k_2 passages from D_2 such that $k_1 + k_2 = k$. Here k_1 and k_2 are the same across different queries. This approach takes into account the fact that D_1 and D_2 are drawn from different distributions and the model being better calibrated on one, handles retrieval from D_1 and D_2 differently.
- Confidence-based allocation: Retrieve some number of passages from D_1 till the cumulative probability of the retrieved passages exceeds some threshold p[2]. The remainder of the budget is allocated to D_2. Unlike the previous approach this one is query-adaptive. This method was inspired by that of nucleus sampling (Holtzman et al., 2020).
- Oracular: Divide the budget k is a way that gives the best retrieval. This is also done per query.

We found fractional allocation to give up to 8 points improvement in recall over Naïve merging.

5 Experiments

Recall from the section on training that we had two corpora D_1 (seen) and D_2 (unseen), and a dataset Q_1 of query-relevant passage tuples over D_1. We also have test queries for which one passage needs to be retrieved from D_1 and D_2 each.

5.1 Dataset Details

One of our primary contributions is the creation of benchmarks for our novel multi-distribution retrieval task owing to the lack of existing ones. We adapted datasets from prior work on different tasks like question answering and entity matching in order to create our benchmarks. Particularly, we use a modified version of CONCURRENTQA (Arora et al., 2022) which is a dataset constructed for investigating privacy-preserving QA. It consists of multi-hop questions over Wikipedia and Enron emails corpora (forming two distinct distributions). We use a subset of questions (called comparison questions[3]) which can be answered in a single-hop and which require passages to be retrieved from both the corpora. There are 100 such question in the validation and test sets each. The Wikipedia and emails corpora have 5.2M and 47k passages respectively. There are roughly 4000 question which require retrieval from one corpus only and we use these for training our models.

The second dataset is created using entity matching datasets from Das et al. (2017). We use the Walmart-Amazon dataset where given a product title, the goal is to retrieve its description from both sources. We provide the scatter plot of BERT embeddings of the product descriptions in Figure 5 to show that they represent different distributions. The Amazon and Walmart corpora have 21891 and 2520 products respectively. We also have a mapping between the listing of the same item in the two site for 1127 items which we use for evaluation (split between validation and test sets in a 1:1 ratio). There are actually 2×1227 queries since product title of either source can be used as the query. The unmatched products are used for training the retrievers. When Amazon is assumed to be the known distribution, we train on (title, description) tuples from Amazon and evaluate on the matched items.

Finally, the third dataset is based on the Amazon-GoogleProducts entity matching dataset from Köpcke et al. (2010) which is similar to the Walmart-Amazon dataset above. The Amazon and Google corpora have 1248 and 3035 products respectively and a mapping between 1161 products is also present. All dataset statistics are summarized in Table 2.

[2]In practice, we retrieve some number $N \gg k$ passages and compute probability as the softmax of the similarity scores. This is needed since the corpus D_i can be very large making computing the softmax infeasible.

[3]eg. "Do ShopBack and Cognitive Arts both deal with internet based services?"

Dataset	$\lvert D_1 \rvert$	$\lvert D_2 \rvert$	Num val	Num test
CONCURRENTQA	5.2M	47k	100	100
Walmart-Amazon	2520	21891	1127	1127
Amazon-Google	1248	3035	1161	1161

Table 2: Number of passages in each corpus and the number of queries for the three datasets used.

5.2 Training Details

We use a shared RoBERTa base encoder as the query and passage encoder. We fine-tune it on Q_1 for 50 epochs using Adam optimizer with a learning rate of $5e-5$ with warmup for 10% of the iterations and a batch size of 64 on four TITAN V GPUs. Training take between 4-10 hours depending on the dataset. Further, we truncate queries to 70 tokens and passages to 300. We also present results on fine-tuning `all-MiniLM-L6-v2` model from `sentence-transformers` (Reimers and Gurevych, 2019) which is a model specifically trained to generate embeddings of sentences and paragraphs for clustering and semantic search.

5.3 Evaluation

Given a query q, we compute its embedding $E_Q(q)$ find the dot product similarity with the embeddings of passages in D_1 and D_2. We then retrieve some number of documents with the highest similarity scores depending on the allocation strategy to get k candidate passages $\mathcal{C}_q$. Finally, the Recall@k is reported. Recall is the number of correct passages retrieved. In our experiments, for a single query, it is 1 if both passages are retrieved, 0.5 if one is retrieved and 0 otherwise. Further, for user facing applications, the rank at which a result is presented is also important. This aspect is not captured by Recall and hence we also report the MRR (mean reciprocal rank). Given a list of ranked items, the MRR is the reciprocal of the rank at which the true passage appears in the rank list i.e. $\mathrm{MRR} = 1/\mathrm{rank}(\text{true passage},\ \text{rank list})$.

5.4 Results

The results on the three datasets is shown in Table 3. The Recall@k and MRR values are reported. We choose k=10 for Walmart-Amazon and Amazon-Google datasets, and k=100 for CONCURRENTQA due its much larger corpus size. The effect of changing k is investigated in the Analysis section.

	Walmart Amazon (k=10)		Amazon Google (k=10)		ConcurrentQA (k=100)	
Known distribution →	Walmart	Amazon	Amazon	Google	Wikipedia	Enron*
Naïve Merging	58.71/0.49	75.26/0.64	67.76/0.55	70.26/0.57	52.00/0.10	80.00/0.26
Oracle	68.29/0.51	85.89/0.69	73.19/0.56	74.70/0.58	61.00/0.10	88.00/0.29
Fractional 0.0	41.23/0.33	45.35/0.39	44.05/0.35	44.40/0.35	51.00/0.09	74.00/0.18
Fractional 0.1	55.63/0.47	69.51/0.63	61.25/0.52	64.14/0.55	55.00/0.10	85.00/0.29
Fractional 0.2	58.54/0.49	76.66/0.67	65.09/0.55	68.75/0.57	54.00/0.10	84.00/0.29
Fractional 0.3	59.93/0.50	78.80/0.68	67.11/0.55	70.26/0.58	54.00/0.10	87.00/0.29
Fractional 0.4	61.61/0.50	80.31/0.68	68.19/0.56	71.21/0.58	55.00/0.10	88.00/0.29
Fractional 0.5	**62.02/0.50**	**80.72/0.68**	**68.84/0.56**	**71.47/0.58**	56.00/0.10	**88.00/0.29**
Fractional 0.6	61.44/0.50	80.49/0.68	68.62/0.56	70.65/0.58	**56.00/0.10**	87.00/0.29
Fractional 0.7	61.15/0.50	80.37/0.68	67.54/0.55	69.22/0.57	53.00/0.10	87.00/0.29
Fractional 0.8	56.10/0.47	76.13/0.66	59.05/0.51	59.40/0.53	48.00/0.10	83.00/0.29
Fractional 0.9	28.11/0.21	41.11/0.31	30.17/0.24	31.08/0.26	46.00/0.10	80.00/0.29
Fractional 1.0	28.11/0.21	41.11/0.31	30.17/0.24	31.08/0.26	17.00/0.02	53.00/0.11
Confidence 0.9	56.10/0.47	76.13/0.66	59.05/0.51	59.40/0.53	-	-

Table 3: Various allocation strategies evaluated on Walmart-Amazon and CONCURRENTQA. The reported numbers in each cell are Recall@k and MRR. The chosen k is 10 for Walmart-Amazon, Amazon-Google and 100 for CONCURRENTQA.

We can see that fractional allocation performs better than the naïve baseline across all datasets achieving upto 8 points higher recall. As one would expect, performance improves with increasing allocation fraction up to a point before decreasing. Both extremes are not good since the retrieval budget gets allocated to a single corpus. It is interesting to note that the fractions close to 0.5 work

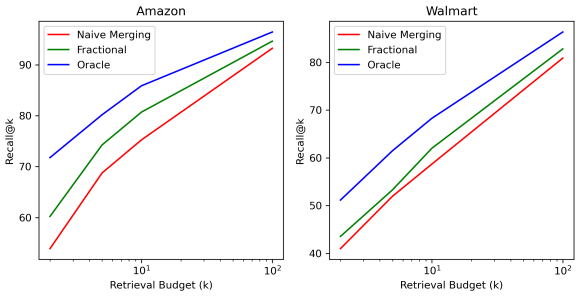

Figure 2: Effect of increasing retrieval budget k. With increasing k, difference between the three methods decreases while the relative ordering remains the same.

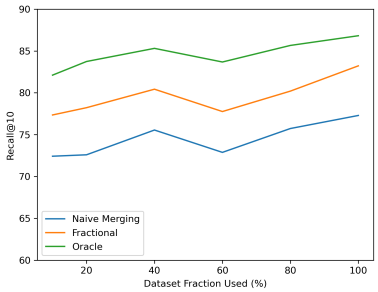

Figure 3: Effect of training data size. Recall increases with training set size but gap between methods remains similar.

best across different settings even though the corpus sizes are imbalanced. Further, on some datasets, the best fractional allocation achieves recall close to that of the oracle while on some there is a gap of 5+ points. We hope that future research on this task will help close this gap.

It was challenging to get confidence based allocation to work well since the distribution of similarity scores were very peaked leading to large peaks in the softmax probabilities. Changing the softmax temperature did not help. However, for completeness, one set of numbers for this method is also reported. We can see that is almost always performs worse than the baseline.

6 Analysis

In this section, we evaluate the effect of various hyperparameters and design choices on the performance of our method. First, let us look at the **effect of increasing the retrieval budget** k. As we can see in Figure 2, recall increases as we increase k for all the three methods. The fractional allocation numbers plotted here are for the best fraction for that choice of k. An interesting observation is that the difference between the three methods decreases as we increase k. This is probably due to the fact that when a large budget is available, the correct passage will be retrieved somewhere in the ranked list of passages even though it might have a somewhat lower similarity score.

Next, we investigate the **effect of training data size** on the performance of the three methods in Figure 3. We plot the Recall@10 values against increasing fractions of the training set being used for Walmart-Amazon. The model used here was `all-MiniLM-L6-v2` since it was smaller than RoBERTa and hence faster to train. As one would expect, the recall increases as more data becomes available but the gains are quite modest. This indicates that even small amount of training data from the known distribution $\mathcal{D}_1$ should be enough to fine-tune the encoders.[4]

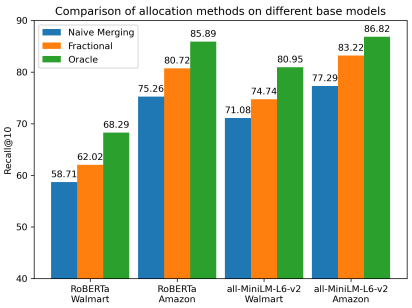

Figure 4: Effect of choice of pretrained model. `all-MiniLM-L6-v2` performs better than RoBERTa but relative ordering between methods remains the same.

Finally, we evaluate the **effect of the choice of pretrained model** by checking if the observations made when fine-tuning RoBERTa also hold for other models. For this, we fine-tune the `all-MiniLM-L6-v2` model on the Walmart-Amazon dataset and compare the performance of the baseline, the best fractional allocation and the oracle in Figure 4. Being trained to generate sentence embeddings for semantic search, `all-MiniLM-L6-v2` performs better than RoBERTa but the relative ordering between the three methods remains the same indicating that our observations can hold across different models.

[4]Though not related to our task, it might be interesting to evaluate the implications of having a few samples from $\mathcal{D}_2$ since it is appears that fine-tuning the encoders does not require a lot of data.

7 Conclusion

In this project, we propose and formalize a novel information retrieval task where the data is drawn from several different distributions, some of which are unknown during training. We created benchmarks for this task and evaluated several simple retrieval methods on these benchmarks. We show that these simple methods work well obtaining up to 8 points improvement in recall over baselines. However, there is a gap of 5+ points between these methods and the oracle indicating avenues for further research. Limitations of our work include the relatively small size of some of the datasets making fine-tuning and evaluation challenging. Future work can try to create larger datasets particularly suited for this task instead of adapting existing datasets from other tasks. More powerful allocation methods can also be developed. Finally, we had restricted to two distributions. Future work can try extending it to multiple distributions.

References

Simran Arora, Patrick Lewis, Angela Fan, Jacob Kahn, and Christopher Ré. 2022. Reasoning over public and private data in retrieval-based systems. *arXiv preprint arXiv:2203.11027*.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine learning*, 79:151–175.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Sanjib Das, AnHai Doan, Paul Suganthan G. C., Chaitanya Gokhale, Pradap Konda, Yash Govind, and Derek Paulsen. 2017. The magellan data repository. `https://sites.google.com/site/anhaidgroup/projects/data`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Hanna Köpcke, Andreas Thor, and Erhard Rahm. 2010. Evaluation of entity resolution approaches on real-world match problems. *Proceedings of the VLDB Endowment*, 3(1-2):484–493.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Hong Liu, Mingsheng Long, Jianmin Wang, and Michael Jordan. 2019a. Transferable adversarial training: A general approach to adapting deep classifiers. In *International Conference on Machine Learning*, pages 4013–4022. PMLR.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jan Nehring, Nils Feldhus, Harleen Kaur, and Akhyar Ahmed. 2021. Combining open domain question answering with a task-oriented dialog system. In *Proceedings of the 1st Workshop on Document-grounded Dialogue and Conversational Question Answering (DialDoc 2021)*, pages 38–45, Online. Association for Computational Linguistics.

Richard Yuanzhe Pang, Alicia Parrish, Nitish Joshi, Nikita Nangia, Jason Phang, Angelica Chen, Vishakh Padmakumar, Johnny Ma, Jana Thompson, He He, and Samuel Bowman. 2022. QuALITY: Question answering with long input texts, yes! In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5336–5358, Seattle, United States. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. 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.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *International Conference on Learning Representations*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Ellen M Voorhees et al. 1999. The trec-8 question answering track report. In *Trec*, volume 99, pages 77–82.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Wenhan Xiong, Xiang Lorraine Li, Srinivasan Iyer, Jingfei Du, Patrick Lewis, William Yang Wang, Yashar Mehdad, Wen-tau Yih, Sebastian Riedel, Douwe Kiela, and Barlas Oğuz. 2021. Answering complex open-domain questions with multi-hop dense retrieval. *International Conference on Learning Representations*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Michael Zhang and Eunsol Choi. 2021. SituatedQA: Incorporating extra-linguistic contexts into QA. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7371–7387, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Appendix

We would like our dataset to satisfy the two properties to be useful for the multi distribution retrieval task. First, we require the presence of two distinct distributions. That in the retrieval corpora D_1 and D_2 should have distinct distributional properties. We plot the t-SNE of BERT embeddings of the passages in the Walmart-Amazon corpus in the Figure 5 to verify this. Further we can see that the queries plotted in the right form a single distribution. Second, we require that the retrieval from the chosen corpora is non-trivial. We quantify this as retrieval from chosen corpora requiring some in-domain training to achieve reasonable performance. We plot the distribution of zero-shot retrieval ranks using BERT embeddings on the Walmart-Amazon dataset to show this.

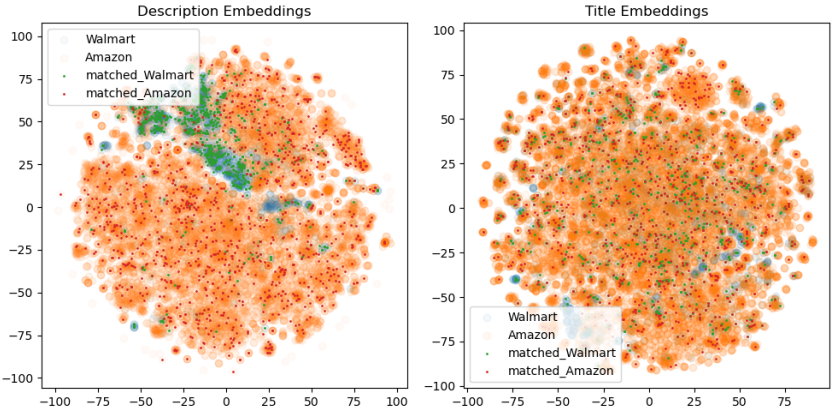

Figure 5: t-SNE plot of the BERT embeddings for the Walmart-Amazon product descriptions and title. It can be seen that the descriptions form two separate distributions while the titles do not.

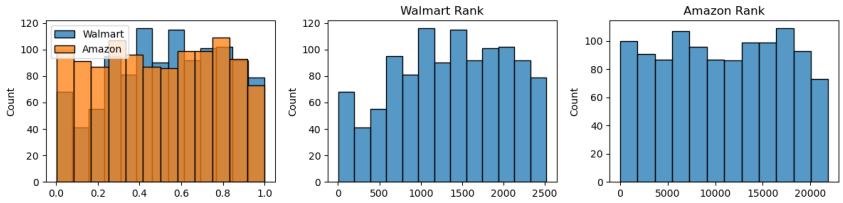

Figure 6: Distributions of ranks of zero-shot retrieval using BERT embeddings. The left figure is the normalized rank while the other two have the actual ranks.