Evaluating Niche Domains with Edu-MiniMARCO

Samuel Gorman
Stanford University
sgorman@stanford.edu

Abstract

Information retrieval in natural language understanding (NLU) has benefitted from the increased effectiveness of large language models with transformer-based architectures such as BERT and T5. Accepted paradigms in the NLU community suggest that to increase model performance, increase the amount of training data that a large-language model may access. Niche domains such as academia, healthcare, law and government often require more precise results with less room for error. In these fields, large amounts of relevant textual data may be significantly more difficult to obtain, prompting the need to investigate solutions that achieve higher accuracy with less data. We present Edu-MiniMARCO, a novel dataset comprised of academic natural language queries from MSMARCO passage-ranking with a size 9,600 training set and 1,700 dev/ test set. We demonstrate that all tested passage reranking approaches achieve lower mean reciprocal rank (MRR) on Edu-MiniMARCO compared to MSMARCO. Empirically, we find that models trained on MSMARCO outperform models trained on Edu-MiniMARCO when tested on Edu-MiniMARCO, affirming the correlation between more data and improved performance even in niche domains.

1 Introduction

In recent years, the information retrieval community has benefitted from large datasets such as TREC-CAR (Dietz et al., 2017) and MSMARCO passage-ranking (Nguyen et al., 2016). For example, MSMARCO consists of 1 million real natural language queries and approximately 9 million passages.

Large datasets have often heralded large increases in model performance for subtasks, with examples such as ImageNet (Deng et al., 2009) for computer vision and SQuAD 2.0 (Rajpurkar and Liang, 2018) for question-answering.

Large language models built with transformers (Vaswani et al., 2017) such as BERT (Devlin et al., 2019) have been successfully adapted to IR tasks, with reranker BERT (Nogueira and Cho, 2019), ColBERT (Khattab and Zaharia, 2020), DocT5query (Nogueira et al., 2019) and TCT-ColBERT (Lin et al., 2020) demonstrating strong gains in accuracy as a reranker to rankings created by sparse vector representations like BM25 (Beaulieu et al., 1997).

However, despite the progress enabled by these large datasets, narrow fields such as education, medicine and law remain skeptical of being able to meaningfully apply deep learning to their industries (Surden, 2019) (Holzinger et al., 2019) (Holmes et al., 2019). Here, we adopt the term niche domain to refer to these fields. More formally, we offer a definition of niche domain as a field where (1) labels are necessary and expensive to obtain (2) relatively low-amounts of quality, labeled training data exist (3) tolerance for error is low in real-life contexts. These characteristics make niche domains a challenging arena for NLU models, where models must perform inference with high accuracy and train with smaller amounts of in-domain data.

In this paper, we focus on the niche domain of education, applied to the task of passage reranking. This task accepts as input a textual query, often in the form of a natural-language question, and outputs the top K ranking of relevant documents. Our education subset, titled Edu-MiniMARCO, is built from education-related queries in MSMARCO and is approximately $\frac{1}{100}$ of the size of MSMARCO.

We adopt mean reciprocal rank (MRR) as our primary evaluation metric, where Q represents a sample of queries and and $rank_i$ represents the ranked position of the first document that is relevant

for the i-th query.

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i} \quad (1)$$

Our contributions to the research community are as follows.

(1) We introduce Edu-MiniMARCO, a novel dataset built from a small amount of education-related queries from MSMARCO to encourage exploration of model performance on niche domains.

(2) We demonstrate that baseline neural models achieve lower MRR at various Ks on Edu-MiniMARCO than MSMARCO.

(3) We show that training on MSMARCO leads to a two-fold increase in MRR compared to training on Edu-MiniMARCO when evaluating on Edu-MiniMARCO.

2 Related Works

Our work is in the domain of neural information retrieval, with a particular focus on the task of passage reranking. Here, we survey the rapid advances made in information retrieval with the introduction of large datasets and neural model architectures, and discuss the efficiency-effectiveness tradeoff present in these models.

2.1 Neural Information Retrieval

Classical methods in IR such as the probabilistic model BM25 found it challenging to achieve higher MRRs, as they often struggled to capture contextual relationships between words when deciding relevance. Earlier papers such as Guo et al in 2016 argued that IR tasks required models distinct from semantic matching NLP tasks, and proposed fine-tuned deep relevance matching models based on histogram matching patterns. (Guo et al., 2016) However, the largest breakthroughs occurred with deep neural architectures.

Early neural ranking models such as KNRM and Duet rely on bespoke model architectures that do not generalize outside of IR tasks (Xiong et al., 2017)(Mitra et al., 2017). Large gains in MRR were demonstrated in 2019 when applying the general-purpose large language model BERT to passage reranking tasks (Nogueira and Cho, 2019). Further research showed that BERT is broadly effective across retrieval tasks (Dai and Callan, 2019). Both authors observed that BERT embeddings outperformed classical bag-of-words approaches, achieving MRRs on MSMARCO over 15 percentage points higher than BM25. However, these models introduce tradeoffs in efficiency as discussed below.

2.2 Efficiency-effectiveness tradeoff

The field of IR has struggled to balance the efficiency-effectiveness tradeoff when constructing model architectures. When training on large corpora such as MS MARCO, average latency per query has been linearly correlated with effectiveness in MRR. Dai demonstrated that increased accuracy was possible with the introduction of BERT embeddings, but this came at the cost of increased query latency (Dai and Callan, 2019). Further, this demands a query-document paradigm, where BERT relies on reranking pipelines from a reranker such as BM25. Khattab and Zaharia introduced ColBERT to provide end-to-end reranking and represent documents as matrices rather than vectors in order to search much faster and directly (Khattab and Zaharia, 2020). This helped to eliminate some of the redundant calculations of BERT. Later work (Hofstätter et al., 2020) has built upon the efficiency-effectiveness question by distiling knowledge from larger BERT models to several efficient models in IR, and (Sanh et al., 2019) introduced DistilBERT, a BERT model that is 40 percent smaller and retains 97 percent of BERT's language-understanding capacities. We use ColBERT as our reranker in our experimental condition for its speed relative to competing BERT models and accuracy. Khattab and Zaharia demonstrated that ColBERT is up to 170× faster and requires 14k× less FLOPs/query than alternative BERT models.

2.3 Large Datasets

Benchmark-driven datasets have accelerated the advancement of neural models across deep learning communities. Examples outside of IR first helped advance fields such as computer vision (Deng et al., 2009). Within IR, large datasets like MSMARCO and TRECCAR have provided direction to the research community (Nguyen et al., 2016) (Dietz et al., 2017).

In a break from benchmark-driven datasets, increases in performance of transformer-based large language models (LLMs) has mainly scaled linearly with the amount of data they are trained on, with models such as T5 and GPT-3 trained with billions of parameters from the open-web (Brown et al., 2020)(Raffel et al., 2020).

Researchers often take these pretrained models and finetune them for their specific, downstream tasks to attempt to increase performance.

However, is it necessary to finetune pretrained large language models on in-domain data to achieve higher performance in niche domains? Task-specific datasets in niche domains such as education and law are often small in size and expensive to produce labels for (Hendrycks et al., 2021). This led us to investigate whether LLMs trained on large amounts of readily-available, general data can achieve parity or exceed performance of LLMs trained on in-domain data.

3 MSMARCO and Edu-MiniMARCO

MSMARCO (Microsoft Machine Reading Comprehension) is a large-scale dataset introduced in 2016 comprised of 1 million real natural language queries and approximately 9 million passages. We are interested in the passage reranking task of the dataset, in which given a query and the K most relevant passages, a system must rerank the most relevant passages as high as possible. Queries are anonymized and aggregated from the search engine Bing, and passages are real snippets taken from 3.5 million web urls. On average, human annotators judged 1.1 passages as relevant to a query, and these relevant passages are represented as `qrels` files in TREC format.

We constructed Edu-MiniMARCO as a subset of academic natural language queries from MSMARCO passage-ranking. Edu-MiniMARCO's defining characteristics are that it (1) strictly contains education-related queries and (2) is $\frac{1}{100}$ of the size of the full MSMARCO corpus. Edu-MiniMARCO consists of a size 9,689 training set, size 900 dev set, and size 900 test set. Each entry takes the form of a `query_id, query pair`. We title the dataset Edu-MiniMARCO to support future MiniMARCO datasets in other niche domains, such as Law-MiniMARCO or Med-MiniMARCO. We sample from the MSMARCO training set to construct the Edu-MiniMARCO training set, and sample from MSMARCO dev set to construct the Edu-MiniMARCO dev/test sets. We first filter MSMARCO to only contain queries that have corresponding `qrels` in order to be able to compute MRR for each query. For example, this reduced the dev set size from 101092 to 55576. We then created a subset of academic natural language queries, where whether a given query was considered academic was determined by whether it (1) contains one or more of 70+ education-related target keywords (2) passes a manual review to filter spurious results. Finally, we drop duplicate queries and removed 311 duplicates in the training set and 200 in the test/dev sets. We have released the dataset publicly on Github to enable future work.

Keyword	**Train** (%)	**Dev / Test** (%)
cells	1502 (0.15)	126 (0.07)
graph	1214 (0.12)	136 (0.07)
atom	772 (0.08)	107 (0.06)
education	731 (0.08)	87 (0.04)
science	717 (0.07)	64 (0.03)
theory	587 (0.06)	51 (0.03)
angle	568 (0.06)	80 (0.04)
math	551 (0.06)	75 (0.04)
history	525 (0.05)	79 (0.04)
biology	513 (0.05)	49 (0.03)

Table 1: **The top 10 keywords out of 70+ contained in size 9,600 train set and size 1,800 dev / test sets. Edu-MiniMARCO. We observe an unequal distribution of keywords where the top 10 keywords represent 0.89 percent of the training set and 0.45 percent of the dev / test sets. Note that these distributions were calculated before dropping duplicate values and are approximations.**

3.1 Observations

Sentences in Edu-MiniMARCO are slightly longer than sentences in MSMARCO. Sentences are a mean length of 6.62 in Edu-MiniMARCO with standard deviation 3.04, while sentences are a mean length of 5.98 with standard deviation 2.43.

The root type-token ratio (RTTR), a measure commonly used to assess textual lexical diversity, rated Edu-MiniMARCO as slightly more complex, with 2.47 and 2.36 for Edu-MiniMARCO and MSMARCO respectively.

As reported in Table 1, the top 10 keywords in the dataset command a dominant distribution compared to the others. Table 2 demonstrates sample natural language queries represented in the dataset.

4 Metrics and Methods

4.1 Metrics

We choose **MRR@10** as our primary metric of evaluation. We conduct further error analysis with Recall@100 and Success@$1, 5, 10, 20, 50, 100$. We analyze Success@ various Ks in order to gain

Sample Queries
”what is the branch of biology that deals with the realations of organisms to one another and to thier physical surroundings”
”phospholipid bilayer definition biology”
”meaning of edges in math”
”what is exponential notation in chemistry”
”what are some examples of anaerobic respiration”

Table 2: **Sample real natural language queries of our education subset from MSMarco.**

insight into the performance of top ranked results that may be difficult to discern from MRR@10. Recall@100 provides insight into how results perform at deeper levels of recall than MRR@10 or Success@K will provide.

4.2 Methods

We empirically evaluate Edu-MiniMARCO by comparing the performance of baseline and neural models against MSMARCO. We then evaluate whether training on the smaller, more specialized Edu-MiniMARCO provides any increases in accuracy compared to training on MSMARCO when evaluating Edu-MiniMARCO.

BM25 We begin with a baseline of BM25, a bag-of-words ranking function that probabilistically ranks sets of documents based on given terms in queries. We compare the BM25 rankings of the Edu-MiniMARCO test set against the BM25 rankings of the MSMARCO dev set. We set K = 1000 to compute the rankings for 1000 relevant documents for each query. Here, we use the MSMARCO dev set as an effective proxy for the test set, which is withheld by the creators of MSMARCO to maintain the integrity of the dataset.

BM25 + ColBERT We chose to adopt ColBERT as a BERT-based reranker. ColBERT's low-latency computations of late interaction over BERT-base leads to dramatically faster training and inference time on retrieval tasks with less compute compared to current large language models. This was an appropriate choice given constraints on access to compute. We feed in earlier BM25 rankings to ColBERT to rerank them. Like in our baseline, we compare the BM25 + ColBERT rankings of the Edu-MiniMARCO test set against the BM25 + ColBERT rankings of the MSMARCO dev set. This provides efficient quantitative results for comparison of MRR across the two datasets.

Training ColBERT on Edu-MiniMARCO We then train ColBERT on the Edu-MiniMARCO training set. We compare this to ColBERT trained on the MSMARCO training set. We then evaluate both trained models on the Edu-MiniMARCO test set. This enables comparison on the effectiveness between training on much smaller, more specialized sets of data, or on large amounts of general data.

4.3 Implementation

We computed BM25 rankings for Edu-MiniMARCO at $K = 1000$, and made use of Pysereni, an open-source Python toolkit built as a wrapper for Anserini (Yang et al., 2018). Anserini exposes the text search engine library Lucerne in Java. We evaluated these rankings against the `qrels` file of passages labeled as relevant using a utility evaluation script provided by ColBERT in order to compute MRR and Recall.

We hosted the open-source ColBERT Github repository on an AWS virtual machine. We then constructed a training file comprised of query, positive passage, negative passage triples from the Edu-MiniMARCO training set, and saved a model checkpoint after training for approximately 600 epochs with a batch size of 16. Finally, we performed inference with this trained model by providing our Edu-MiniMARCO test set, TopK@1000 BM25 rankings, and matching `qrels` file.

4.4 Hardware

We train our ColBERT model on an AWS g4dn.2xlarge virtual machine with a single NVIDIA T4 GPU, 8 virtual CPUs and 32GB of memory.

5 Experimental Evaluation

We now address the following research questions.

Q1: Will baseline and neural models achieve a higher MRR when tested on Edu-MiniMARCO or MSMARCO?

Q2: Will neural models trained on MSMARCO achieve a higher MRR on Edu-MiniMARCO than models trained on Edu-MiniMARCO?

5.1 Testing on Edu-MiniMARCO vs MSMARCO

We compare the results of evaluating the same models on Edu-MiniMARCO and MSMARCO as reported in Table 3. We observe that models evaluated on the Edu-MiniMARCO achieve lower MRR@10 than those same models evaluated on MSMARCO, with an approximate difference in

Method	Tested On	MRR@10	Recall@50	Recall@200	Recall@1000
BM25	MSMARCO	0.187	0.592	0.738	0.857
BM25	Edu-MiniMARCO	0.152	0.526	0.677	0.843
BM25+ ColBERT	MSMARCO	0.348	0.753	0.805	0.814
BM25 + ColBERT	Edu-MiniMARCO	0.299	0.756	0.826	0.843

Table 3: **Reranking results of BM25 baseline and BM25+ ColBERT reranker tested on MSMARCO and Edu-MiniMARCO. We observe lower MRR when testing identical methods on Edu-MiniMARCO.**

Method	Trained On	MRR@10	MRR@100	Recall@50	Recall@200	Recall@1000
ColBERT	MSMARCO	**0.299**	**0.31**	**0.756**	**0.826**	0.843
ColBERT	Edu-MiniMARCO	0.140	0.151	0.527	0.739	0.843

Table 4: **Comparison of results on Edu-MiniMARCO test set. Here, we compare ColBERT trained on the MSMARCO train set to ColBERT trained on the Edu-MiniMARCO train set. We observe significantly higher results when trained on the full MSMARCO train set.**

MRR of 4.5 between the two datasets. This provides quantitative evidence that niche domains may present an increased challenge for model performance compared to general tasks. Note that we provide evidence supporting lower MRR in the niche domain of education, and leave verification of additional niche domains for future work.

5.2 Training on Edu-MiniMARCO vs MSMARCO

We compare the MRR and Recall at various Ks of ColBERT trained on Edu-MiniMARCO compared to ColBERT trained on MSMARCO , as reported in Table 2. We evaluate these models on the Edu-MiniMARCO test set. We find that models trained on the full MSMARCO corpus significantly outperform those trained on the the smaller, more specific Edu-MiniMARCO training set.

On one hand, these results affirm recent work that has shown that larger amounts of training data tends to increase model performance. Nogueira and Cho found a positive correlation between MRR@10 and the total number of examples BERT trained on with MSMARCO (Nogueira and Cho, 2019). Outside of information retrieval, (Raffel et al., 2020) ablated aspects of pretraining such as model-size and cleanness of data and found that only scaling model size and the amount of training data led to significant increases in performance.

On the other hand, these results quantitatively demonstrate that this principle applies to niche domains where availability of relevant training data may be more sparse. We show that training on large amounts of out-of-domain data outperformed smaller in-domain data by a factor of 2.

6 Conclusion and Future Work

In this paper, we introduced Edu-MiniMARCO, a novel dataset built from a small amount of education-related queries from MSMARCO to encourage exploration on model performance when testing on niche domains. We demonstrated that MRR across models decreases on Edu-MiniMARCO compared to MSMARCO, suggesting that niche domains are uniquely more challenging to achieve strong performance on than general-purpose use cases. And when evaluating Edu-MiniMARCO, we showed that training on the larger, more general MSMARCO corpus led to a two-fold increase in MRR compared to training on the smaller, in-domain Edu-MiniMARCO corpus.

We have made the full Edu-MiniMARCO dataset and the code to reproduce our experiments publicly available. We encourage future work in the creation of additional MiniMARCO datasets to evaluate performance in more niche domains such as medicine or law. Our work suggests that these niche domains pose a greater challenge for model performance, and more exploration is needed to determine whether alternatives exist apart from providing models with more training data to address this gap in MRR. We support future work in evaluating these results beyond passage-ranking, such as in question-answering.

7 Acknowledgements and Contributions

We thank Omar Khattab and Professor Christopher Potts for discussing ideas and providing direction to the project as part of Stanford CS224u: Natural Language Understanding.

References

Micheline Beaulieu, Mike Gatford, Xiangji Huang, S Robertson, Steve Walker, and P Williams. 1997. Okapi at trec-5. *Nist Special Publication SP*, pages 143–166.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Zhuyun Dai and Jamie Callan. 2019. Deeper text understanding for ir with contextual neural language modeling. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 985–988.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

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.

Laura Dietz, Manisha Verma, Filip Radlinski, and Nick Craswell. 2017. Trec complex answer retrieval overview. In *TREC*.

Jiafeng Guo, Yixing Fan, Qingyao Ai, and W Bruce Croft. 2016. A deep relevance matching model for ad-hoc retrieval. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, pages 55–64.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. Cuad: An expert-annotated nlp dataset for legal contract review. *arXiv preprint arXiv:2103.06268*.

Sebastian Hofstätter, Sophia Althammer, Michael Schröder, Mete Sertkan, and Allan Hanbury. 2020. Improving efficient neural ranking models with cross-architecture knowledge distillation. *arXiv preprint arXiv:2010.02666*.

Wayne Holmes, Maya Bialik, and Charles Fadel. 2019. Artificial intelligence in education. *Boston: Center for Curriculum Redesign*.

Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. 2019. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4):e1312.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 39–48.

Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2020. Distilling dense representations for ranking using tightly-coupled teachers. *arXiv preprint arXiv:2010.11386*.

Bhaskar Mitra, Fernando Diaz, and Nick Craswell. 2017. Learning to match using local and distributed representations of text for web search. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1291–1299.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. In *CoCo@ NIPS*.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*.

Rodrigo Nogueira, Jimmy Lin, and AI Epistemic. 2019. From doc2query to doctttttquery. *Online preprint*.

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.

Robin Jia Rajpurkar, Pranav and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Harry Surden. 2019. Artificial intelligence and law: An overview. *Georgia State University Law Review*, 35:19–22.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Chenyan Xiong, Zhuyun Dai, Jamie Callan, Zhiyuan Liu, and Russell Power. 2017. End-to-end neural ad-hoc ranking with kernel pooling. In *Proceedings of the 40th International ACM SIGIR conference on research and development in information retrieval*, pages 55–64.

Peilin Yang, Hui Fang, and Jimmy Lin. 2018. Anserini: Reproducible ranking baselines using lucene. *Journal of Data and Information Quality (JDIQ)*, 10(4):1–20.