

Synthetic Target Domain Supervision for Open Retrieval QA

Revanth Gangi Reddy
revanth3@illinois.edu
UIUC
United States

Bhavani Iyer
bsiyer@us.ibm.com
IBM Research AI
United States

Md Arafat Sultan
arafat.sultan@ibm.com
IBM Research AI
United States

Rong Zhang
zhangr@us.ibm.com
IBM Research AI
United States

Avirup Sil
avi@us.ibm.com
IBM Research AI
United States

Vittorio Castelli
vittorio@us.ibm.com
IBM Research AI
United States

Radu Florian
raduf@us.ibm.com
IBM Research AI
United States

Salim Roukos
roukos@us.ibm.com
IBM Research AI
United States

ABSTRACT

Neural passage retrieval is a new and promising approach in open retrieval question answering. In this work, we stress-test the Dense Passage Retriever (DPR)—a state-of-the-art (SOTA) open domain neural retrieval model—on closed and specialized target domains such as COVID-19, and find that it lags behind standard BM25 in this important real-world setting. To make DPR more robust under domain shift, we explore its fine-tuning with synthetic training examples, which we generate from unlabeled target domain text using a text-to-text generator. In our experiments, this noisy but fully automated target domain supervision gives DPR a sizable advantage over BM25 in out-of-domain settings, making it a more viable model in practice. Finally, an ensemble of BM25 and our improved DPR model yields the best results, further pushing the SOTA for open retrieval QA on multiple out-of-domain test sets.

CCS CONCEPTS

• Information systems → Question answering; • Computing methodologies → Natural language generation.

KEYWORDS

Open retrieval question answering, Neural passage retrieval, Weak supervision, Out-of-domain neural IR

ACM Reference Format:

Revanth Gangi Reddy, Bhavani Iyer, Md Arafat Sultan, Rong Zhang, Avirup Sil, Vittorio Castelli, Radu Florian, and Salim Roukos. 2021. Synthetic Target Domain Supervision for Open Retrieval QA. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)*, July 11–15, 2021, Virtual Event, Canada. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3404835.3463085>

1 INTRODUCTION

Open retrieval question answering (ORQA) finds a short answer to a natural language question in a large document collection [4, 9, 26]. Most ORQA systems employ (i) an information retrieval

(IR) component that retrieves relevant passages from the given corpus [14, 26, 35] and (ii) a machine reading comprehension (MRC) component that extracts the final short answer from a retrieved passage [2, 29, 33]. Recent work on ORQA by Karpukhin et al. [21] shows that distant supervision for neural passage retrieval can be derived from annotated MRC data, yielding a superior approach [17, 28] to classical term matching methods like BM25 [9, 34]. Concurrent advances in tools like FAISS [19] that support efficient similarity search in dense vector spaces have also made this approach practical: when queried on an index with 21 million passages, FAISS processes 995 questions per second (qps). BM25 processes 23.7 qps per CPU thread in a similar setting [21].

Crucially, all training and test instances for the Dense Passage Retrieval (DPR) model in [21] were derived from open domain Wikipedia articles. This is a rather limited experimental setting, as many real-world ORQA use cases involve distant target domains with highly specialized content and terminology, for which there is no labeled data. On COVID-19, for example, a large body of scientific text is available [38], but practically no annotated QA data for model supervision.¹ In this paper, we closely examine neural IR—DPR to be specific—in out-of-domain ORQA settings, where we find that its advantage over BM25 diminishes or disappears altogether in the absence of target domain supervision.

Domain adaptation is an active area of investigation in supervised learning; existing techniques for different target scenarios include instance weighting [18], training data selection using reinforcement learning [30] and transfer learning from open domain datasets [39]. For pre-trained language models, fine-tuning on unlabeled target domain text has also been found to be a useful intermediate step [13, 40], e.g., with scientific [7] and biomedical text [3, 25]. To address the performance degradation of DPR in low-resource out-of-domain settings, we explore another approach: fine-tuning with synthetically generated examples in the target domain. Our example generator is trained using open domain (Wikipedia) MRC examples [33]. It is then applied to target domain (biomedical) documents to generate synthetic training data for both retrieval and MRC. Despite being trained on generic open domain annotations, our generator yields target domain examples that significantly boost results in those domains. It should be noted here that unlike most existing work in the QA literature where human annotated training examples are used to fine-tune a synthetically

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](https://permissions.acm.org).
SIGIR '21, July 11–15, 2021, Virtual Event, Canada

© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8037-9/21/07...\$15.00
<https://doi.org/10.1145/3404835.3463085>

¹The handful of existing COVID-19 QA datasets [24, 32, 37] are quite small in size and can only be used for evaluation.

Passage	Synthetic Question-Answer pairs
... Since December 2019, when the first patient with a confirmed case of COVID-19 was reported in Wuhan, China, over 1,000,000 patients with confirmed cases have been reported worldwide. It has been reported that the most common symptoms include fever, fatigue, dry cough, anorexia, and dyspnea. Meanwhile, less common symptoms are nasal congestion ...	Q: What are the most common symptoms of COVID-19? A: fever, fatigue, dry cough, anorexia, and dyspnea Q: How many people have been diagnosed with COVID-19? A: over 1,000,000

Table 1: Synthetic MRC examples generated by our generator from a snippet in the CORD-19 collection.

pre-trained model [1, 11, 12, 36], we rely on only synthetic examples in the target domain.

The contributions of this paper are as follows:

- We empirically show the limitations of open domain neural IR (DPR) when applied zero shot to ORQA in distant target domains.
- We present a solution to this problem that relies on automatic text-to-text generation to create target domain synthetic training data. Our synthetic examples improve *both IR and end-to-end ORQA results*, in *both original and related target domains*, requiring *no supervision with human annotated examples*.
- We also show that ensembling over BM25 and our improved neural IR model yields the best results—which underscores the complementary nature of the two approaches—further pushing the state of the art for out-of-domain ORQA on multiple benchmarks.

2 METHOD

This section describes our methods for generating synthetic examples in the target domain and their application to both IR and MRC to construct the final ORQA pipeline.

2.1 Generating Synthetic Training Examples

Let (p, q, a) be an MRC example comprising a passage p , a question q , and its short answer a in p . Let s be the sentence in p that contains the answer a . In what follows, we train an example generator to produce the triple (s, a, q) given p . The answer sentence s is subsequently used to locate a in p , as a short answer text (*e.g.*, a named entity) can generally occur more than once in a passage.

To train the generator, we fine-tune BART [27]—a pre-trained denoising sequence-to-sequence generation model—with MRC examples from open domain datasets like SQuAD [33]. The generator g with parameters θ_g learns to maximize the conditional joint probability $P(s, a, q|p; \theta_g)$. In practice, we (i) only output the first (s_f) and the last (s_l) word of s instead of the entire sentence for efficiency, and (ii) use special separator tokens to mark the three items in the generated triple.

Given a target domain passage p at inference time, an ordered sequence $(s_f, s_l, [SEP], a, [SEP], q)$ is sampled from g using top- k top- p sampling [15], which has been shown to yield better training examples than greedy or beam search decoding due to greater sample diversity [36]. From this generated sequence, we create positive synthetic training examples for both passage retrieval: (q, p) and MRC: (p, q, a) , where s_f and s_l are used to locate a in p . Table 1 shows two examples generated by our generator from a passage in the CORD-19 collection [38].

2.2 Passage Retrieval

As stated before, we use DPR [21] as our base retrieval model. While other competitive methods such as ColBERT [22] exist, DPR offers a number of advantages in real-time settings as well as in our target scenario where retrieval is only a component in a larger ORQA pipeline. For example, by compressing each passage down to a single vector representation, DPR can operate with significantly less memory. It is also a faster model for several reasons, including not having a separate re-ranking module.

For target domain supervision of DPR, we fine-tune its off-the-shelf open domain instance with synthetic examples. At each iteration, a set of questions is randomly sampled from the generated dataset. Following Karpukhin et al. [21], we also use in-batch negatives for training. We refer the reader to their article for details on DPR supervision. We call this final model the *Adapted DPR* model.

2.3 Machine Reading Comprehension

For MRC, we adopt the now standard approach of Devlin et al. [10] that (i) starts from a pre-trained transformer language model (LM), (ii) adds two pointer networks atop the final transformer layer to predict the start and end positions of the answer phrase, and (iii) fine-tunes the entire network with annotated MRC examples. We choose RoBERTa [31] as our base LM. Given our out-of-domain target setting, we fine-tune it in two stages as follows.

First, the RoBERTa LM is fine-tuned on unlabeled target domain documents, which is known to be a useful intermediate fine-tuning step [13]. This target domain model is then further fine-tuned for MRC, where we use both human annotated open domain MRC examples and target domain synthetic examples, as detailed in Section 3. Additionally, we denoise the synthetic training examples using a roundtrip consistency [1] filter: an example is filtered out if its candidate answer score, obtained using an MRC model trained on SQuAD 2.0 and NQ, is lower than a threshold t (t tuned on a validation set).

2.4 Open Retrieval Question Answering

Using the described retrieval and MRC components, we construct our final ORQA system that executes a four-step process at inference time. First, only the K highest scoring passages returned by IR for the input question are retained (K tuned on a validation set). Each passage is then passed along with the question to the MRC component, which returns the respective top answer and its MRC score. At this point, each answer has two scores associated with it: its MRC score and the IR score of its passage. In the third step, these two scores get normalized using the Frobenius norm and combined using a convex combination. The weight in the combination operation is tuned on a validation set. Finally, the answer with the highest combined score is returned.

Model	Open-COVID-QA-2019						COVID-QA-111		
	Dev			Test			Test		
	M@20	M@40	M@100	M@20	M@40	M@100	M@20	M@40	M@100
BM25	22.4	24.9	29.9	29.9	33.4	39.7	48.7	60.4	64.9
DPR-Multi	14.4	18.4	22.9	13.8	17.5	21.4	51.4	57.7	66.7
ICT	16.6	21.6	25.5	18.1	23.0	29.6	52.8	59.8	67.6
Adapted DPR	28.0	31.8	39.0	34.8	40.4	47.2	58.6	64.6	74.2
BM25 + DPR-Multi	23.4	27.9	32.3	29.5	33.2	38.9	58.6	65.8	69.4
BM25 + Adapted DPR	31.8	36.0	42.6	43.2	48.2	53.7	60.4	68.2	76.9

Table 2: Performance of different IR systems on (a) the open retrieval version of COVID-QA-2019, and (b) COVID-QA-111.

3 EXPERIMENTAL SETUP

We evaluate the proposed systems on out-of-domain retrieval, MRC, and end-to-end ORQA against SOTA open domain baselines.

3.1 Retrieval Corpus and Datasets

We select COVID-19 as our primary target domain, an area of critical interest at the point of the writing. We use 74,059 full text PDFs from the June 22, 2020 version of CORD-19 [38] document collection on SARS-CoV-2—and related coronaviruses as our retrieval corpus. Each document is split into passages that (a) contain no more than 120 words, and (b) align with sentence boundaries, yielding around 3.5 million passages.

We utilize three existing datasets for COVID-19 target domain evaluation. The first one, used to evaluate retrieval and MRC results separately as well as end-to-end ORQA performance, is *COVID-QA-2019* [32]—a dataset of question-passage-answer triples created from COVID-19 scientific articles by volunteer biomedical experts. We split the examples into Dev and Test subsets of 203 and 1,816, respectively. Since end-to-end ORQA examples consist of only question-answer pairs with no passage alignments, we also create a version of this dataset for ORQA evaluation (*Open-COVID-QA-2019* henceforth) wherein duplicate questions are de-duplicated and different answers to the same question are all included in the set of correct answers, leaving 201 Dev and 1,775 Test examples.

Our second dataset—*COVID-QA-147* [37]—is a QA dataset obtained from Kaggle’s CORD-19 challenge, containing 147 question-article-answer triples with 27 unique questions and 104 unique articles. Due to the small number of unique questions in this dataset, we only use it for out-of-domain MRC evaluation.

Finally, *COVID-QA-111* [24] contains queries gathered from different sources, e.g., Kaggle and the FAQ sections of the CDC and the WHO. It has 111 question-answer pairs with 53 interrogative and 58 keyword-style queries. Since questions are not aligned to passages in this dataset, we use it only to evaluate IR and ORQA.

3.2 Synthetic Example Generation

We fine-tune BART for three epochs on the open domain MRC training examples of SQuAD1.1 [33] ($\text{lr}=3\text{e-}5$). Synthetic training examples are then generated for COVID-19 from the CORD-19 collection. We split the articles into chunks of at most 288 wordpieces and generate five MRC examples from each of the resulting 1.8 million passages. For top- k top- p sampling, we use $k=10$ and $p=0.95$. Overall, the model generates about 7.9 million examples.

3.3 Retrieval and MRC

We use the DPR-Multi system from [21] as our primary neural IR baseline. DPR-Multi comes pre-trained on open-retrieval versions of several MRC datasets: Natural Questions (NQ) [23], WebQuestions [8], CuratedTrec [6] and TriviaQA [20]. We fine-tune it for six epochs with COVID-19 synthetic examples to train our *Adapted DPR* model ($\text{lr}=1\text{e-}5$, batch size=128). We also evaluate the Inverse Cloze Task (ICT) method [26] as a second neural baseline, which masks out a sentence at random from a passage and uses it as a query to create a query-passage training pair. We use ICT to fine-tune DPR-Multi on the CORD-19 passages of Section 3.2, which makes it also a synthetic domain adaptation baseline. Finally, for each neural IR model, we also evaluate its ensemble with BM25 that computes a convex combination of normalized neural and BM25 scores. The weight for BM25 in this combination is 0.3 (tuned on Open-COVID-QA-2019 Dev).

Our baseline MRC model is based on a pre-trained RoBERTa-Large LM, and is fine-tuned for three epochs on SQuAD2.0 and then for one epoch on NQ. It achieves a short answer EM of 59.4 on the NQ dev set, which is competitive with numbers reported in [29]. For target domain training, we first fine-tune a RoBERTa-Large LM on approximately 1.5GB of CORD-19 text containing 225 million tokens (for 8 epochs, $\text{lr}=1.5\text{e-}4$). The resulting model is then fine-tuned for MRC for three epochs on SQuAD2.0 examples and one epoch each on roundtrip-consistent synthetic MRC examples and NQ. For roundtrip consistency check, we use a threshold of $t=7.0$, which leaves around 380k synthetic examples after filtering.

3.4 Metrics

We evaluate IR using Match@ k , for $k \in \{20, 40, 100\}$ [21]. For MRC, we use standard Exact Match (EM) and F1 score. Finally, end-to-end ORQA accuracy is measured using Top-1 and Top-5 F1.

4 RESULTS AND ANALYSIS

We first report results separately for IR and MRC. Then we evaluate ORQA pipelines that must find a short answer to the input question in the CORD-19 collection. Reported numbers for all trained models are averages over three random seeds.

4.1 Passage Retrieval

Table 2 shows performances of different IR systems on Open-COVID-QA-2019 and COVID-QA-111. BM25² demonstrates robust results

²Lucene Implementation. BM25 parameters $b = 0.75$ (document length normalization) and $k_1 = 1.2$ (term frequency scaling) worked best.

Model	Open-COVID-QA-2019				COVID-QA-111	
	Dev		Test		Test	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
BM25 → Baseline MRC	21.7	31.8	27.1	38.7	24.1	39.3
(BM25 + DPR-Multi) → Baseline MRC	21.4	30.9	25.2	37.2	24.4	43.2
(BM25 + Adapted DPR) → Baseline MRC	24.2	35.6	29.5	44.2	25.0	45.9
(BM25 + Adapted DPR) → Adapted MRC	27.2	37.2	30.4	44.9	26.5	47.8

Table 3: End-to-end F1 scores achieved by different Open retrieval QA systems. The best system (last row) utilizes target domain synthetic training examples for both IR and MRC supervision.

relative to the neural baselines. While DPR-multi is competitive with BM25 on COVID-QA-111, it is considerably behind on the larger Open-COVID-QA-2019. ICT improves over DPR-multi, indicating that even weak target domain supervision is useful. The proposed Adapted DPR system achieves the best single system results on both datasets, with more than 100% improvement over DPR-Multi on the Open-COVID-QA-2019 Test set. Finally, ensembling over BM25 and neural approaches yields the best results. The BM25+Adapted DPR ensemble is the top system across the board, with a difference of at least 14 points with the best baseline on the Open-COVID-QA-2019 Test set (all metrics), and 8 points on COVID-QA-111.

Upon closer examination, we find that BM25 and Adapted DPR retrieve passages that are very different. For Open-COVID-QA-2019, for example, only 5 passages are in common on average between the top 100 retrieved by the two systems. This diversity in retrieval results explains why they complement each other well in an ensemble system, leading to improved IR performance.

4.2 Machine Reading Comprehension

Table 4 shows results on the two COVID-19 MRC datasets. Input to each model is a question and an annotated document that contains an answer. Our proposed model achieves 2.0–3.7 F1 improvements on the Test sets over a SOTA open domain MRC baseline. On the COVID-QA-2019 Dev set, we see incremental gains from applying the two domain adaptation strategies.

Model	COVID-QA-2019				COVID-QA-147	
	Dev		Test		Test	
	EM	F1	EM	F1	EM	F1
Baseline MRC	34.0	59.4	34.7	62.7	8.8	31.0
+ CORD-19 LM	35.5	60.2	-	-	-	-
+ Syn. MRC training	38.6	62.8	37.2	64.7	11.3	34.7

Table 4: MRC performances on COVID-19 datasets. The last row refers to the proposed model that is trained on unlabeled CORD-19 text as well as synthetic MRC examples.

4.3 Open Retrieval Question Answering

Using different pairings of the above IR and MRC systems, we build several ORQA pipelines. Each computes a convex combination of its component IR and MRC scores after normalization, with the IR weight being 0.7 (tuned on Open-COVID-QA-2019 Dev). We observe that retrieving $K=100$ passages is optimal when IR is BM25 only, while $K=40$ works best for BM25+Neural IR.

Table 3 shows end-to-end F1 scores of the different ORQA pipelines. Adapted MRC refers to the best system of Section 4.2 (Table 4 Row

3). Crucially, the best system in Table 3 (last row) uses synthetic target domain supervision for both IR and MRC. In a paired t -test [16] of the Top-5 F1 scores, we find the differences with the baseline (Row 1) to be statistically significant at $p < 0.01$.

4.4 Zero Shot Evaluation on BioASQ

To investigate if our synthetically fine-tuned COVID-19 models can also help improve performance in a related target domain, we evaluate them zero shot on the BioASQ [5] task. BioASQ contains biomedical questions with answers in the PubMed abstracts. For evaluation, we use the factoid questions from the Task 8b training and test sets, totaling 1,092 test questions. As our retrieval corpus, we use around 15M abstracts from Task 8a. We pre-process them as described in Section 3.1 to end up with around 37.4M passages.

Model	M@20	M@40	M@100
BM25	42.1	46.4	50.5
DPR-Multi	37.6	42.8	48.1
Adapted DPR	42.4	48.9	55.9

Table 5: IR results on BioASQ Task 8B factoid questions.

Table 5 shows the BioASQ retrieval results, where the proposed Adapted DPR model outperforms both baselines. Table 6 summarizes the evaluation on the end-to-end ORQA task, where we see similar gains from synthetic training. These results show that synthetic training on the CORD-19 articles transfers well to the broader related domain of biomedical QA.

Model	Top-1	Top-5
BM25 → Baseline MRC	30.6	45.5
DPR-Multi → Baseline MRC	28.6	43.0
Adapted DPR → Baseline MRC	32.1	49.4
Adapted DPR → Adapted MRC	32.9	49.5

Table 6: ORQA F1 scores on BioASQ 8B factoid questions.

5 CONCLUSION

Low-resource target domains can present significant challenges for supervised language processing systems. In this paper, we show that synthetically generated target domain examples can support strong domain adaptation of neural open domain open retrieval QA models, which can further generalize to related target domains. Crucially, we assume zero labeled data in the target domain and rely only on open domain MRC annotations to train our generator. Future work will explore semi-supervised and active learning approaches to examine if further improvements are possible with a small amount of target domain annotations.

REFERENCES

- [1] Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. Synthetic QA Corpora Generation with Roundtrip Consistency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 6168–6173.
- [2] Chris Alberti, Kenton Lee, and Michael Collins. 2019. A BERT baseline for the Natural Questions. *arXiv preprint arXiv:1901.08634* (2019).
- [3] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly Available Clinical BERT Embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. 72–78.
- [4] Akari Asai, Jungo Kasai, Jonathan H. Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2020. XOR QA: Cross-lingual Open-Retrieval Question Answering. *arXiv:2010.11856* [cs.CL]
- [5] Georgios Balikas, Anastasia Krithara, Ioannis Partalas, and George Paliouras. 2015. BioASQ: A Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering. In *Revised Selected Papers from the First International Workshop on Multimodal Retrieval in the Medical Domain-Volume 9059*. 26–39.
- [6] Petr Baudiš and Jan Šedivý. 2015. Modeling of the question answering task in the yodaqa system. In *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 222–228.
- [7] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 3606–3611.
- [8] Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 conference on empirical methods in natural language processing*. 1533–1544.
- [9] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to Answer Open-Domain Questions. In *Association for Computational Linguistics (ACL)*.
- [10] 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)*. 4171–4186.
- [11] Bhuwan Dhingra, Danish Danish, and Dheeraj Rajagopal. 2018. Simple and Effective Semi-Supervised Question Answering. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. 582–587.
- [12] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified Language Model Pre-training for Natural Language Understanding and Generation. In *Proceedings of NeurIPS*.
- [13] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don’t Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 8342–8360.
- [14] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909* (2020).
- [15] Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. 2020. The Curious Case of Neural Text Degeneration. In *ICLR*.
- [16] Henry Hsu and Peter A Lachenbruch. 2005. Paired t test. *Encyclopedia of Biostatistics* 6 (2005).
- [17] Gautier Izacard and Edouard Grave. 2020. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. *arXiv preprint arXiv:2007.01282* (2020).
- [18] Jing Jiang and ChengXiang Zhai. 2007. Instance weighting for domain adaptation in NLP. In *Proceedings of the 45th annual meeting of the association of computational linguistics*. 264–271.
- [19] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data* (2019).
- [20] Mandar Joshi, Eunsol Choi, Daniel S 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)*. 1601–1611.
- [21] Vladimir Karpukhin, Barlas Ögüz, Sewon Min, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. *arXiv preprint arXiv:2004.04906* (2020).
- [22] 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*. 39–48.
- [23] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics* 7 (2019), 453–466.
- [24] Jinhyuk Lee, Sean S Yi, Minbyul Jeong, Mujeen Sung, Wonjin Yoon, Yonghwa Choi, Miyoung Ko, and Jaewoo Kang. 2020. Answering Questions on COVID-19 in Real-Time. *arXiv preprint arXiv:2006.15830* (2020).
- [25] J Lee, W Yoon, S Kim, D Kim, S Kim, CH So, and J Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics (Oxford, England)* 36, 4 (2020), 1234.
- [26] Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent Retrieval for Weakly Supervised Open Domain Question Answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 6086–6096.
- [27] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 7871–7880.
- [28] Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2005.11401* (2020).
- [29] Dayiheng Liu, Yeyun Gong, Jie Fu, Yu Yan, Jiusheng Chen, Daxin Jiang, Jiancheng Lv, and Nan Duan. 2020. RikiNet: Reading Wikipedia Pages for Natural Question Answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 6762–6771.
- [30] Miaofeng Liu, Yan Song, Hongbin Zou, and Tong Zhang. 2019. Reinforced training data selection for domain adaptation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 1957–1968.
- [31] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).
- [32] Timo Möller, G Anthony Reina, Raghavan Jayakumar, and Lawrence Livermore. 2020. COVID-QA: A Question Answering Dataset for COVID-19. (2020).
- [33] 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*. 2383–2392.
- [34] Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval* 3, 4 (2009), 333–389.
- [35] Minjoon Seo, Jinhyuk Lee, Tom Kwiatkowski, Ankur Parikh, Ali Farhadi, and Hannaneh Hajishirzi. 2019. Real-Time Open-Domain Question Answering with Dense-Sparse Phrase Index. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 4430–4441.
- [36] Md Arafat Sultan, Shubham Chandel, Ramón Fernandez Astudillo, and Vittorio Castelli. 2020. On the Importance of Diversity in Question Generation for QA. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 5651–5656.
- [37] Raphael Tang, Rodrigo Nogueira, Edwin Zhang, Nikhil Gupta, Phuong Cam, Kyunghyun Cho, and Jimmy Lin. 2020. Rapidly Bootstrapping a Question Answering Dataset for COVID-19. *arXiv preprint arXiv:2004.11339* (2020).
- [38] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, et al. 2020. CORD-19: The COVID-19 Open Research Dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.
- [39] Georg Wiese, Dirk Weissenborn, and Mariana Neves. 2017. Neural Domain Adaptation for Biomedical Question Answering. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*. 281–289.
- [40] Rong Zhang, Revanth Gangi Reddy, Md Arafat Sultan, Vittorio Castelli, Anthony Ferritto, Radu Florian, Efsun Sarioglu Kayi, Salim Roukos, Avirup Sil, and Todd Ward. 2020. Multi-Stage Pretraining for Low-Resource Domain Adaptation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 5461–5468.