Embedding-based Zero-shot Retrieval through Query Generation

Davis Liang ♠,†   **Peng Xu** ♠,†   **Siamak Shakeri** ◇*   **Cicero Nogueira dos Santos** ♠
{liadavis, pengx, cicnog}@amazon.com, siamaks@google.com

Ramesh Nallapati♠   **Zhiheng Huang**♠   **Bing Xiang**♠
{rnallapa, zhiheng, bxiang}@amazon.com

AWS AI Labs♠, Google◇

Abstract

Passage retrieval addresses the problem of locating relevant passages, usually from a large corpus, given a query. In practice, lexical term-matching algorithms like BM25 are popular choices for retrieval owing to their efficiency. However, term-based matching algorithms often miss relevant passages that have no lexical overlap with the query and cannot be finetuned to downstream datasets. In this work, we consider the embedding-based two-tower architecture as our neural retrieval model. Since labeled data can be scarce and because neural retrieval models require vast amounts of data to train, we propose a novel method for generating synthetic training data for retrieval. Our system produces remarkable results, significantly outperforming BM25 on 5 out of 6 datasets tested, by an average of 2.45 points for Recall@1. In some cases, our model trained on synthetic data can even outperform the same model trained on real data.

1 Introduction

We consider the large-scale ad-hoc retrieval problem—retrieving relevant documents from a large corpus, given a query. Algorithms such as BM25 (Robertson and Zaragoza, 2009), based on lexical term-matching, are among the most enduring and well-weathered models in classic information retrieval (IR). In fact, they enjoy widespread use in state-of-the-art ranking systems, where typically a deep neural ranking model re-ranks the BM25 retrieval results (Nogueira and Cho, 2019; MacAvaney et al., 2019; Yilmaz et al., 2019; Nogueira et al., 2020).

However, lexical matching algorithms are unable to capture semantic similarities not involving lexical overlap, are not trainable on target datasets, and
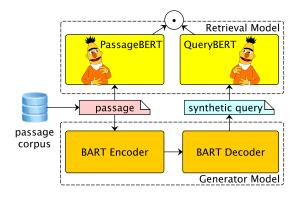
---

*Work done at Amazon.
† Corresponding Authors

Figure 1: Synthetic queries, generated by a sequence-to-sequence model, are used to train our two-tower retrieval architecture.

cannot fully leverage recent advances in pretrained representations (Devlin et al., 2018; Liu et al., 2019). These limitations make term-matching algorithms sub-optimal for passage ranking or question answering (Lee et al., 2019; Guu et al., 2020; Karpukhin et al., 2020).

In this work, we consider the embedding-based two-tower architecture for neural retrieval. In this model, query and document embeddings are constructed from two encoders (i.e., "towers") and the documents are ranked based on the similarity of these embeddings. Since document embeddings can be pre-computed, online retrieval is efficient with various approximate nearest-neighbor search algorithms (Aumüller et al., 2017). The two-tower architecture, originated from the Siamese network (Bromley et al., 1994), is particularly well-studied for the retrieval task (Severyn and Moschitti, 2015; dos Santos et al., 2015; Das et al., 2016; Ma et al., 2019; Cer et al., 2018; Reimers and Gurevych, 2019; Lee et al., 2019; Chang et al., 2020).

As with all deep neural models, the two-tower architecture relies on the availability of a large amount of training data. However, for retrieval,

Original Passage	Synthetic Queries
Medical errors affect one in 10 patients worldwide . One extrapolation suggests that 180,000 people die each year partly as a result of iatrogenic injury . One in five Americans (22 %) report that they or a family member have experienced a medical error of some kind . The World Health Organization registered 14 million new cases and 8.2 million cancer-related deaths in 2012 . It estimated that the number of cases could increase by 70 % through 2032 . As the number of cancer patients receiving treatment increases , hospitals around the world are seeking ways to improve patient safety , to emphasize traceability and raise efficiency in their cancer treatment processes .	how many cancer deaths are preventable how many people die from cancer every year number of deaths from cancer due to medical errors how many deaths are caused by miscommunication how many americans a year experience a medical error
For the rest of her life , she kept a cloth tied to her eyes and thus deprived herself of the power of sight . At certain critical junctures , she gave advice to her husband which was impeccable from a moral standpoint ; she never wavered in her adherence to dharma (righteousness) , even to a very bitter end . She was fated to witness the death of all her hundred sons within the space of 18 days , during the Great War between them and their cousins ; she also curses the lord Krishna when she was full of sorrow on the death of her 100 children that his vansh (Clan) would also be destroyed in the same manner as that of her .	meaning of dharma is dharma righteous who said, dharma is righteousness why was vedanta so bitter why is there a garment tied to your eyes

Table 1: Examples of synthetic queries from WIKIGQ.

obtaining labeled training data can often be expensive or even impossible. One possible solution exists with weak supervision; one can utilize noisy but cheap supervision signals such as query logs (MacAvaney et al., 2017), traditional IR results (Dehghani et al., 2017), or an ensemble of multiple sources (Xu et al., 2019) to create training data. However, even these weakly supervised approaches rely on either an existing query set or external query log files, which are not always available.

In this paper, we leverage synthetic queries generated from a large sequence-to-sequence (seq2seq) model for pretraining and unsupervised domain adaptation of the two-tower model. Specifically, we finetune BART (Lewis et al., 2019a) on MS-MARCO positive query-passage pairs to perform query generation (QG). Then, we construct a large-scale dataset by applying the BART QG model on English Wikipedia passages to generate synthetic query-passages pairs. We find that the BART-based QG model can generate surprisingly high-quality synthetic queries. For some datasets such as NATURAL QUESTIONS, training on these synthetic datasets can surpass the performance of training on the official training set.

In the zero-shot setting, described in Section 5, the Siamese model pretrained on synthetic Wikipedia queries significantly outperforms BM25 baselines on several datasets from both Wikipedia and non-Wikipedia domains. In addition, we apply our QG model on target domain datasets (e.g., AN-TIQUE, BIOASQ, INSURANCEQA, etc.) to generate domain-specific synthetic data. We demonstrate that finetuning our retrieval models on these domain-specific synthetic queries can further improve performance.

Our contributions are as follows:

- We demonstrate a query generation method that can synthesize large amounts of high-quality data for the retrieval task. Finetuning on *synthetic* queries generated from the target domain to perform *unsupervised domain adaptation* can improve zero-shot performance on those target datasets.
- We achieve state-of-the-art performance on the ReQA benchmark.
- We explore a variety of ablations, considering variations in embedding sizes, architecture choices, decoding methods, and more.

2 Models

In this section, we describe our methodology for query generation and its application to the passage retrieval task as illustrated in Fig. 1.

2.1 Query Generator

We formulate query generation as a seq2seq task, where the input is a passage and the target output is a relevant query. Positive query-passage pairs, used to finetune our query generator, can be found in passage ranking datasets such as MSMARCO (Nguyen et al., 2016) or extractive QA datasets such as SQUAD (Rajpurkar et al., 2016). After finetuning the query generator, we apply it to per-

Dataset	Domain	# Passages	# Evaluation queries	# Train queries	Length of passages mean (std)
WIKIGQ	Wiki	22M	-	-	123.3 (32.5)
NATURAL QUESTIONS	Wiki	84,783	4,340	110,865	128.2 (74.2)
TREC-CAR	Wiki	3.5M	195,659	584,461	92.9 (77.5)
MSMARCO	non-Wiki	8.8M	55,578	532,761	77.0 (30.9)
INSURANCEQA	non-Wiki	27,288	1,625	10,391	121.3 (82.5)
ANTIQUE	non-Wiki	403,666	200	2,426	54.3 (74.8)
BIOASQ	non-Wiki	20M	500	2,747	288.7 (130.6)
REQA SQUAD	Wiki	97,707	87,599	-	209.4 (79.7)
REQA NQ	Wiki	239,013	74,097	-	193.2 (112.2)

Table 2: Summary of the datasets.

form query synthesis on arbitrary text corpora. We consider two use cases. First, we apply our generator to passages from English Wikipedia to create a large-scale synthetic pretraining dataset. Second, we apply the generator to the passages of target domain corpora and synthesize queries to create domain-specific synthetic data. Neither case requires human-annotated data in the target domain.

In this work, we adopt BART (Lewis et al., 2019a), a pretrained Transformer-based seq2seq model, as our query generator.

2.2 Passage Retriever

BERT (Devlin et al., 2018) is a bidirectional transformer encoder model pretrained on English Wikipedia and BooksCorpus (Zhu et al., 2015). We leverage BERT-base as the foundation for our retrieval models. On top of BERT-base, we extract the `[CLS]` embedding, followed by a dense layer with a `tanh` activation function, to obtain the final query/passage representation.

- **Two-Tower** retrieval models consist of two *independent* BERT-base models: one for each of the query and passage towers, respectively. During inference, we retrieve passages whose embeddings are the most similar to the query embedding.

- **Siamese** retrieval models additionally require that the passage and query BERT models share the same parameters. Siamese networks (Bromley et al., 1994) are popular among tasks that involve quantifying the similarity between comparable items, here a query and a passage.

Training Objective For our retrieval model, we leverage the dot product between query and passage embeddings as a measure of relevance. Following Lee et al. (2019), we use sampled Softmax

during training.

3 Datasets

In this section, we describe all the datasets that are used in our experiments. A summary of the datasets is shown in Table 2.

WikiGQ (Wikipedia Generated Queries) is the dataset synthesized by our QG model. It consists of 110M synthetically generated queries on 22M passages from our English Wikipedia dataset[1]. See Table 1 for some examples. This dataset will be open-sourced for research purposes.

Natural Questions (Kwiatkowski et al., 2019) is an end-to-end question answering dataset constructed from English Wikipedia. We convert NATURAL QUESTIONS into a passage retrieval dataset, keeping queries that have long answers. We discard answers that are not regular paragraphs (e.g., tables, lists). Ultimately, we retain 110,865 of 307,372 queries from the official training set and 4,340 of 7,842 queries from the official development set. Collecting long answers from these queries nets us 84,783 passages. We hold out the 4,340 development queries as the test set.

TREC-CAR (Dietz et al., 2017) is a dataset created from English Wikipedia for complex answer retrieval, where the queries are generated by concatenating articles and section titles and the ground truth passages consist of the paragraphs within that section. We use the last segment from the predefined 5 segments of the dataset as our test set.

MSMARCO (Nguyen et al., 2016) comprises over 1M queries sampled from Bing search query logs. Specifically, we use the MSMARCO Passage

---

[1]We take the official Wikipedia 2016 database dump and split individual Wikipedia articles into passages with maximum 100 words each, respecting sentence boundaries.

Ranking (PR) dataset, which contains 8.8M passages from over 3.5M web documents retrieved by Bing. We use the official development set of 55,578 queries as our test set. We also use the official training set of 532,761 queries and their positive passages to train our query generation model.

InsuranceQA(v2) (Feng et al., 2015) consists of 2,000 evaluation queries on passages sourced from the Insurance Library database.

ANTIQUE (Hashemi et al., 2020) is a collection of 2,626 open-domain, non-factoid questions sourced from Yahoo! Answers. We use the provided test set of 200 queries with the corpus of over 400K passages for evaluation.

BioASQ (Tsatsaronis et al., 2012) Task B consists of 500 English evaluation queries and reference answers constructed by a team of biomedical experts (BioASQ 7b). We collect approximately 20M non-empty article abstracts sourced from the PubMed database as our passage corpus.

ReQA (Ahmad et al., 2019) is a benchmark for evaluating sentence-level answer retrieval models. This benchmark consists of two datasets, REQA SQUAD and REQA NQ, created from the official training sets of SQuAD and Natural Questions, respectively. Each candidate passage is a sentence concatenated with its context passage and a relevant candidate is the answer sentence concatenated with its context passage.

4 Implementation Details

Generation We begin by finetuning the pretrained BART-large (374M parameters) on MS-MARCO, generating queries given a passage. We train BART for 5 epochs with a batch size of 96 and a learning rate of 3e-5 (warmup ratio 0.1). Once the BART model is trained, we perform generation over English Wikipedia passages to generate 10 synthetic queries for each passage, using nucleus sampling (Holtzman et al., 2020) with $p = 0.95$, and retaining the top-5 queries based on likelihood scores. The resulting dataset consists of 110M synthetic query-passage pairs, which we call WIKIGQ.

Pretraining We initialize our Siamese/two-tower retrieval models with BERT-base (110M parameters). We then pretrain our retrieval models on WIKIGQ. The pretraining is done on 32 Nvidia V100 GPUs with a batch size of 100 per GPU for 10 epochs. We set a max learning rate of 5e-5 with

warmup ratio of 0.1. The pretraining takes about 4 days for the Siamese model and 6 days for the Two-Tower model.

For our baseline models trained on the MS-MARCO official training set, we train for 10 epochs on 8 GPUs with a batch size of 50 per GPU and max learning rate of 3e-5 with warmup ratio of 0.1. Half-precision training[2] is enabled in all of our experiments.

For our Inverse Cloze Task (ICT) baselines, we follow the process defined in Lee et al. (2019). For each passage, we randomly sample one sentence as a query and remove it from the passage 90% of the time. To ensure fair comparison, we train on the ICT data, where data is generated on-the-fly, for 5 times as many epochs as on the QG synthetic data in order to match the number of training iterations. Specifically, we train our models on QG synthetic data for 10 epochs (50 epochs for ICT) on 8 GPUs with a batch size of 50 per GPU and learning rate 3e-5.

Finetuning We use the same BART QG model (trained on MSMARCO) to generate domain-specific synthetic data (we do this for several datasets that are not derived from English Wikipedia: INSURANCEQA, ANTIQUE, BIOASQ). When finetuning our retrieval models on domain-specific synthetic data, we set max learning rate to 3e-5 with warmup ratio of 0.1 and finetune on 8 GPUs with a batch size of 50 per GPU for 10 epochs (5 epochs for BIOASQ).

5 Experimental Results

In this section, we present our main results on zero-shot retrieval. Specifically, we consider the following models and baselines:

- **BM25**: a classical IR method based on lexical term-matching. We consider BM25 as a strong baseline for an unsupervised retrieval system. We leverage the implementation from Elasticsearch[3] with the default settings.
- **Siamese/Two-Tower + MSMARCO**: Because we use MSMARCO to train our BART query generator, we consider a Siamese/Two-Tower model trained on MSMARCO as a reasonable baseline to compare against.
- **Siamese/Two-Tower + WikiGQ**: We train both the Siamese and Two-Tower models on

---

Model	Training data	NATURAL QUESTIONS			TREC-CAR			MSMARCO		
		R@1	R@10	R@100	R@1	R@10	R@100	R@1	R@10	R@100
BM25	None	30.67	63.75	82.02	11.41	32.45	48.86	9.89	36.50	62.83
Two-Tower	MSMARCO	33.56	70.10	87.65	13.45	30.04	44.71	*16.31*	*48.55*	*75.92*
Two-Tower	WIKIGQ	44.33	82.12	94.54	**18.13**	**42.46**	58.49	13.38	49.49	79.93
Siamese	MSMARCO	38.52	73.17	88.36	16.85	36.41	50.88	*18.18*	*53.33*	*80.46*
Siamese	WIKIGQ	**45.07**	**82.91**	**95.32**	14.91	40.20	**58.87**	12.98	50.74	**81.38**
Siamese	OFFICIAL DATA	*40.78*	*80.77*	*94.96*	*22.10*	*48.18*	*67.47*	*18.18*	*53.55*	*80.46*

Model	Training data	INSURANCEQA			ANTIQUE			BIOASQ		
		R@1	R@10	R@100	R@1	R@10	R@100	R@1	R@10	R@100
BM25	None	22.41	50.61	78.98	3.25	17.59	42.16	**20.28**	**45.32**	**71.83**
Two-Tower	MSMARCO	19.74	48.21	75.56	4.79	16.23	36.20	3.25	10.53	22.34
Two-Tower	WIKIGQ	20.27	50.29	78.71	5.15	22.65	46.51	9.22	24.24	43.50
Two-Tower	WIKIGQ+FT(GQ)	30.20	67.27	91.37	5.90	22.45	45.83	14.39	33.46	55.45
Siamese	MSMARCO	23.19	52.52	79.72	5.05	19.04	40.75	6.41	14.64	31.79
Siamese	WIKIGQ	23.24	54.65	81.76	4.15	17.35	43.06	12.30	30.60	52.17
Siamese	WIKIGQ+FT(GQ)	**31.47**	**68.85**	**92.05**	**6.09**	**23.38**	**49.01**	15.91	37.30	61.77
Siamese	OFFICIAL DATA	*30.82*	*67.73*	*92.88*	*3.89*	*18.54*	*41.83*	*6.50*	*19.77*	*36.90*

Table 3: Detailed results on zero-shot performance. Rows with 'MSMARCO' as 'training data' use the official MSMARCO PR training set to train the models. Rows with 'WIKIGQ' use our synthetic WIKIGQ dataset as training data. Rows with 'WIKIGQ+FT(GQ)' denotes further finetuning on domain-specific synthetic data after pretraining on WIKIGQ. Models trained on official training sets, including rows corresponding to OFFICIAL DATA, are in *italics*. Best zero-shot numbers are in **bold**. Numbers are in percent (%).

Model	Training data	Average Recall		
		R@1	R@10	R@100
BM25	None	16.32	41.04	64.45
Two-Tower	MSMARCO	15.18	37.28	57.06
Two-Tower	WIKIGQ	18.41	45.21	66.95
Siamese	MSMARCO	18.03	41.52	61.99
Siamese	WIKIGQ	**18.77**	**46.08**	**68.76**

Table 4: Average results on zero-shot performance. Numbers are in percent(%).

the WIKIGQ dataset.

- **Siamese/Two-Tower + WikiGQ + domain-specific synthetic data**: For datasets that are not based on English Wikipedia, we finetune on the synthetic data additionally generated from said datasets.

The results are shown in Tables 3 & 4. We find that, in most cases, our models trained on the synthetic WIKIGQ data outperforms BM25 and the models trained on MSMARCO. In particular, for datasets such as NATURAL QUESTIONS and TREC-CAR, which are based on English Wikipedia, the recall from the models trained on WIKIGQ is about 20% - 50% relatively higher than the BM25 baselines. For non-Wikipedia datasets such as INSURANCEQA and ANTIQUE, we obtain a significant boost in performance through unsupervised domain adaptation.

One notable exception is the BIOASQ dataset on which our neural retrieval models fail to outperform the BM25 baseline. One possible reason lies in the construction of the BIOASQ dataset itself.[4] In this dataset, human annotations are built from candidates retrieved via term-matching with optional boosting tags (Tsatsaronis et al., 2012). Furthermore, the annotation depth is relatively shallow (approximately 10-60 articles per query) whereas the total number of articles is around 20M. We believe that this annotation process favors lexical term-matching systems like BM25.

Comparing the results from the Siamese and Two-Tower models, we find that the Siamese model generally outperforms the Two-Tower model, indicating that sharing parameters across encoders is helpful. This is consistent with observations from Das et al. (2016).

Finally, for each dataset, we report the performance of the Siamese model trained with official training sets. The results are reported in *italics*

---

[4] Please see details in http://participants-area.bioasq.org/general_information/Task7b/

in the bottom rows of each section in Table 3. Remarkably, we find that in 4 out of 6 datasets, Siamese models trained purely on synthetic data can already outperform the models trained on the official training sets. In particular, for ANTIQUE and BIOASQ, which have relatively small training sets, Siamese models trained on synthetic data improve Recall@1 by over 50% and 140%, respectively. This further demonstrates that QG can be helpful when only a limited amount of labeled data is available.

6   Ablations

Embedding Size   We consider two variations of retrieval model: the Two-Tower model (without shared weights) and the Siamese model (with shared weights). For each variation we test various embedding sizes for the BERT-base encoders. For simplicity, we perform our ablations on the (non-synthetic) NATURAL QUESTIONS dataset. Fig. 2 shows a monotonic increase in performance as the embedding size grows larger, slowly plateauing around an embedding size of 512.
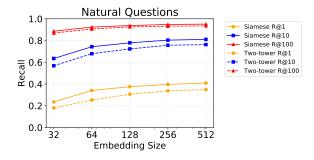

Figure 2: Comparison between Siamese and Two-Tower networks with different embedding sizes.

Generation Method	# pairs	NATURAL QUESTIONS		
		R@1	R@10	R@100
Siamese				
ICT	N/A	18.85	48.26	73.54
Beam	422,309	40.15	77.40	91.51
Nucleus ($p = 0.95$)	363,303	**40.97**	**79.36**	**93.16**
Two-tower				
ICT	N/A	13.72	40.13	67.36
Beam	422,309	35.97	73.19	89.09
Nucleus ($p = 0.95$)	363,303	**37.58**	**75.56**	**91.30**

Table 5: Comparison between ICT and synthetic pre-training. 'Beam' refers to beam search with beam width 5 as the generator decoding strategy. 'Nucleus' refers to taking the top-5 samples from the 10 independent sequences generated via nucleus sampling with $p = 0.95$. Numbers are in percent (%).

Original Query-Passage Pair
Passage: The BMW E70 is the second generation X5 Sports Activity Vehicle (SAV) . It replaced the BMW X5 (E53) in November 2006 . The second generation X5 features many new technological advancements including BMW 's iDrive system as standard equipment and , for the first time in a BMW , an optional third row seat raising passenger capacity to seven .
Original Query: when did the new shape bmw x5 come out
Synthetic Queries (Beam Search):
what is bmw x5 what is bmw e70 what year did bmw x5 come out what year did the bmw x5 come out what year was the bmw x5 made
Synthetic Queries (Nucleus Sampling):
what year did bmw x5 come out what year was the bmw x5 sports activity vehicle what is bmw sav how many people does a bmw x5 seat what year was bmw x5 sports activity vehicle (sav) made

Table 6: Examples of synthetic queries on NATURAL QUESTIONS passages. Synthetic queries from nucleus sampling not only are similar to real queries, but also cover significantly more passage content.

Pretraining Methodology   Recently, several pretraining tasks have been proposed to improve the performance of embedding-based neural retrieval. ICT was first introduced by Lee et al. (2019) with strong performance on open-domain QA tasks. Subsequently, Chang et al. (2020) introduced two new pretraining tasks, Body First Selection (BFS) and Wiki Link Prediction (WLP), in addition to ICT. In this ablation study, we directly compare these pretraining methods against our own.

Beginning with the NATURAL QUESTIONS dataset, we compare models trained with ICT against our own model, trained with synthetic queries. When synthesizing queries, we consider two decoding techniques — beam search and nucleus sampling (Holtzman et al., 2020). With beam search decoding, using a beam width of 5, we generate 5 queries for each passage. Using nucleus sampling, with $p = 0.95$, we generate 10 independent samples, selecting the top 5 queries based on the likelihood scores. After removing duplicate queries, the total number of generated query-passage pairs is shown in Table 5.

We find that both Siamese and Two-Tower models, pretrained with synthetic queries, significantly outperform their ICT counterparts. We hypothesize that models trained on ICT (where the model at-

Model	Pretraining	REQA SQUAD					REQA NQ				
		R@1	R@5	R@10	R@50	R@100	R@1	R@5	R@10	R@50	R@100
USE-QA	-	43.9	65.6	72.7	-	-	14.7	31.7	39.1	-	-
Two-Tower[†]	ICT+BFS+WLP	37.43	61.48	70.18	85.37	89.85	17.31	43.62	55.00	76.59	82.84
BM25	-	**58.51**	**76.80**	**82.14**	90.45	92.72	18.06	42.23	52.26	70.71	76.52
Two-Tower	WIKIGQ	41.29	67.59	76.03	89.10	92.51	21.71	55.55	68.35	86.65	90.83
Siamese	WIKIGQ	46.53	72.52	80.27	**91.42**	**94.19**	**21.88**	**56.48**	**69.67**	**87.93**	**91.88**

Table 7: Zero-short performance on ReQA datasets. USE-QA is reported from (Ahmad et al., 2019) and Two-Tower[†] is reported from (Chang et al., 2020). Since Chang et al. (2020) did not report zero-shot performance, we report their numbers with 1%/99% training/test split setting. Numbers are in percent (%).

tempts to retrieve a passage, given a sentence from that passage) may suffer from this inductive bias when used to perform ranking (where the model must retrieve a passage, given a user query). On the other hand, synthetic queries from our generator are very similar to real user queries. To illustrate this point further, we provide examples in Table 6.

Additionally, we find that training with synthetic queries generated via nucleus sampling outperforms training with queries generated with beam search. While beam search decoding generates strictly more training pairs than nucleus sampling (nucleus sampling can generate duplicates across k independent runs), nucleus sampling tends to result in more diverse queries (Holtzman et al., 2020). We show examples in Table 6.

For completeness, we additionally compare our model against the ICT, BFS and WLP pretraining tasks in conjunction. Following Chang et al. (2020), we use all of English Wikipedia during pretraining for both synthetic query generation (i.e., WIKIGQ as described in Section 4) and evaluate the resulting models on the ReQA benchmark.

We report our zero-shot retrieval results on the BM25 baseline and models trained on WIKIGQ. A summary of the results are shown in Table 7[5]. As we can see, on both datasets, models pretrained on our synthetic WIKIGQ corpus significantly outperform their counterparts pretrained with ICT+BFS+WLP. Notably, the Two-Tower model trained on ICT+BFS+WLP is actually finetuned with 1% of the original training data after ICT+BFS+WLP pretraining. However, our models are trained only on the synthetic WIKIGQ data (no real data was used). These results indicate that

pretraining with WIKIGQ is more favorable than ICT, BFS and WLP for retrieval. Finally, our models pretrained on synthetic data outperform BM25 on REQA NQ by a large margin. However, our model is unable to consistently outperform BM25 on REQA SQUAD. We believe that this is due to certain characteristics of the SQuAD dataset. Similar observations have been made in Lee et al. (2019); Karpukhin et al. (2020)[6].

Model	Recall		
	R@1	R@10	R@100
Siamese	NATURAL QUESTIONS		
FT w/o pretraining	40.78	80.77	94.96
FT w/ WIKIGQ pretraining	**48.57**	**88.29**	**97.30**
Siamese	INSURANCEQA		
FT w/o pretraining	30.82	67.72	92.88
FT w/ WIKIGQ pretraining	**34.33**	**73.46**	**95.71**

Table 8: Abalation study on Siamese models, with and without pretraining on WIKIGQ. Here we use the official training set of NATURAL QUESTIONS and INSURANCEQA for finetuning. Numbers are in percent (%).

With or Without Pretraining We study the impact of WIKIGQ pretraining when we can finetune models on hand-labeled training data. Specifically, we finetune the Siamese models with and without pretrained on WIKIGQ on the official training data of two datasets, NATURAL QUESTIONS and INSURANCEQA. Our results show that pretraining with WIKIGQ improves performance for both Wikipedia-based dataset (NATURAL QUESTIONS) and out-of-domain dataset (INSURANCEQA). We detail our results in Table 8.

---

[5]Our BM25 baseline significantly outperforms those reported in Ahmad et al. (2019); Chang et al. (2020). As described earlier, our BM25 results are obtained from Elasticsearch using default settings. We have communicated with the authors from Ahmad et al. (2019) and they were able to reproduce our results.

[6]They present two possible reasons. First, the high lexical overlap between SQuAD questions and passages due to the fact that the questions were created by the annotators after seeing the passages. Second, data was collected from only 500+ Wikipedia articles which makes the data highly correlated and suboptimal for models trained on i.i.d samples.
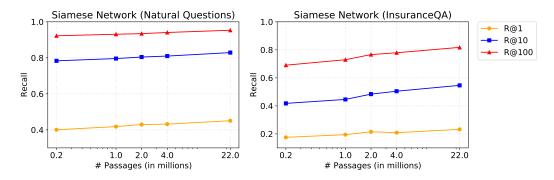
Figure 3: Ablations on amount of passages used for pretraining.

Data Efficiency with Pretraining Finally, we study the sample efficiency of synthetic pretraining with WIKIGQ. We evaluate on two datasets, NATURAL QUESTIONS (Wikipedia-based) and IN-SURANCEQA (non-Wikipedia-based). Specifically, we train the Siamese model on synthetic data generated with various fractions of English Wikipedia. Our results show that performance improves monotonically with synthetic dataset size on both NATURAL QUESTIONS and INSURANCEQA. However, we witness diminishing returns, and training on 22M passages only confers a slight improvement over 4M passages. The results are shown in Fig. 3.

7 Related Work

Recent work has demonstrated the effectiveness of embedding-based neural retrieval models on large document corpora. Lee et al. (2019) first introduced the Inverse Cloze Task (ICT) as a pretraining task for neural retrieval models, demonstrating improved performance over BM25 for open-domain question answering (QA) tasks. Chang et al. (2020) proposed additional pretraining tasks and showed improved performance in the ReQA benchmark (Ahmad et al., 2019). However, none of these works have demonstrated effectiveness in the zero-shot setting.

Along this vein, leveraging generative models to produce synthetic data has been previously explored. Du et al. (2017) first applied the seq2seq model for automatic question generation from text in reading comprehension. Tang et al. (2017); Sachan and Xing (2018) proposed to jointly train QG and QA models to improve QA performance. Lewis et al. (2019b) trained an unsupervised sequence-to-sequence model to generate natural questions from cloze content. Alberti et al. (2019) generated queries and answers by finetuning BERT on extractive subsets of SQuAD 2.0 and

Natural Questions and employing a sequence-to-sequence model for query generation. Puri et al. (2020) first demonstrated that a QA model trained on purely synthetic questions and answers can outperform models trained on human-labeled data on SQuAD1.1. None of these have studied using QG for passage retrieval tasks.

A contemporaneous work uses synthetic queries for domain adaptive neural retrieval (Ma et al., 2020). Our work additionally considers large-scale pretraining with synthetic queries, additional evaluation datasets, and careful ablations.

8 Conclusions

In the paper, we proposed synthetic query generation for improving the performance of embedding-based neural retrieval models in the zero-shot setting. We synthesize WIKIGQ, a large-scale synthetic retrieval dataset from English Wikipedia. Leveraging WIKIGQ, our retrieval models are able to outperform other pretraining strategies such as ICT while also exhibiting superior zero-shot performance on multiple datasets from various domains. Finetuning on the domain-specific synthetic data further improves performance.

Future work on synthetic query generation should address questions about data quality. For example, methods for intelligent filtering to reduce the number of unanswerable queries may lessen the noise present in synthetic data. Decreasing the amount of redundant data through improvements to decoding strategies may further improve training efficiency. Finally, practitioners should consider leveraging better models and larger datasets for training both generation and retrieval models.

References

Amin Ahmad, Noah Constant, Yinfei Yang, and Daniel Cer. 2019. Reqa: An evaluation for end-to-end answer retrieval models. *arXiv preprint arXiv:1907.04780*.

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. Synthetic qa corpora generation with roundtrip consistency. *Association for Computational Linguistics (ACL)*.

Martin Aumüller, Erik Bernhardsson, and Alexander Faithfull. 2017. Ann-benchmarks: A benchmarking tool for approximate nearest neighbor algorithms. In *International Conference on Similarity Search and Applications (SISAP)*, pages 34–49. Springer.

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1994. Signature verification using a" siamese" time delay neural network. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 737–744.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Wei-Cheng Chang, Felix X Yu, Yin-Wen Chang, Yiming Yang, and Sanjiv Kumar. 2020. Pre-training tasks for embedding-based large-scale retrieval. *International Conference on Learning Representations (ICLR)*.

Arpita Das, Harish Yenala, Manoj Chinnakotla, and Manish Shrivastava. 2016. Together we stand: Siamese networks for similar question retrieval. In *Association for Computational Linguistics (ACL)*, pages 378–387.

Mostafa Dehghani, Hamed Zamani, Aliaksei Severyn, Jaap Kamps, and W Bruce Croft. 2017. Neural ranking models with weak supervision. In *Special Interest Group on Information Retrieval (SIGIR)*, pages 65–74.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *North American Chapter of the ACL (NAACL)*.

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

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1342–1352, Vancouver, Canada. Association for Computational Linguistics (ACL).

Minwei Feng, Bing Xiang, Michael R Glass, Lidan Wang, and Bowen Zhou. 2015. Applying deep learning to answer selection: A study and an open task. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 813–820. IEEE.

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*.

Helia Hashemi, Mohammad Aliannejadi, Hamed Zamani, and W Bruce Croft. 2020. Antique: A non-factoid question answering benchmark. In *European Conference on Information Retrieval*, pages 166–173. Springer.

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

Vladimir Karpukhin, Barlas Oğuz, 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*.

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

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. *Association for Computational Linguistics (ACL)*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019a. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Patrick Lewis, Ludovic Denoyer, and Sebastian Riedel. 2019b. Unsupervised question answering by cloze translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4896–4910, Florence, Italy. Association for Computational Linguistics.

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*.

Ji Ma, Ivan Korotkov, Yinfei Yang, Keith Hall, and Ryan McDonald. 2020. Zero-shot neural retrieval via domain-targeted synthetic query generation. *arXiv preprint arXiv:2004.14503*.

Xiaofei Ma, Peng Xu, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2019. Universal text representation from BERT: An empirical study. *arXiv preprint arXiv:1910.07973*.

Sean MacAvaney, Kai Hui, and Andrew Yates. 2017. An approach for weakly-supervised deep information retrieval. *Special Interest Group on Information Retrieval (SIGIR)*.

Sean MacAvaney, Andrew Yates, Arman Cohan, and Nazli Goharian. 2019. Cedr: Contextualized embeddings for document ranking. In *Special Interest Group on Information Retrieval (SIGIR)*, pages 1101–1104.

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.

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

Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. 2020. Document ranking with a pretrained sequence-to-sequence model. *arXiv preprint arXiv:2003.06713*.

Raul Puri, Ryan Spring, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2020. Training question answering models from synthetic data. *arXiv preprint arXiv:2002.09599*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc.

Mrinmaya Sachan and Eric Xing. 2018. Self-training for jointly learning to ask and answer questions. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 629–640.

Cicero dos Santos, Luciano Barbosa, Dasha Bogdanova, and Bianca Zadrozny. 2015. Learning hybrid representations to retrieve semantically equivalent questions. In *Association for Computational Linguistics (ACL)*, pages 694–699.

Aliaksei Severyn and Alessandro Moschitti. 2015. Learning to rank short text pairs with convolutional deep neural networks. In *Special Interest Group on Information Retrieval (SIGIR)*, pages 373–382.

Duyu Tang, Nan Duan, Tao Qin, Zhao Yan, and Ming Zhou. 2017. Question answering and question generation as dual tasks. *arXiv preprint arXiv:1706.02027*.

George Tsatsaronis, Michael Schroeder, Georgios Paliouras, Yannis Almirantis, Ion Androutsopoulos, Eric Gaussier, Patrick Gallinari, Thierry Artieres, Michael R Alvers, Matthias Zschunke, et al. 2012. Bioasq: A challenge on large-scale biomedical semantic indexing and question answering. In *2012 AAAI Fall Symposium Series*.

Peng Xu, Xiaofei Ma, Ramesh Nallapati, and Bing Xiang. 2019. Passage ranking with weak supervision. *ICLR Limited Labeled Data Workshop*.

Zeynep Akkalyoncu Yilmaz, Shengjin Wang, Wei Yang, Haotian Zhang, and Jimmy Lin. 2019. Applying BERT to document retrieval with birch. In *Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *The IEEE International Conference on Computer Vision (ICCV)*.