Towards Robust Neural Retrieval Models with Synthetic Pre-Training

Revanth Gangi Reddy[1], Vikas Yadav[2], Md Arafat Sultan[2], Martin Franz[2],
Vittorio Castelli[2], Heng Ji[1], Avirup Sil[2]
[1]University of Illinois at Urbana-Champaign [2]IBM Research AI, New York
{revanth3,hengji}@illinois.edu, arafat.sultan@ibm.com,
{vikasy,franzm,vittorio,avi}@us.ibm.com

Abstract

Recent work has shown that commonly available machine reading comprehension (MRC) datasets can be used to train high-performance neural information retrieval (IR) systems. However, the evaluation of neural IR has so far been limited to standard supervised learning settings, where they have outperformed traditional term matching baselines. We conduct in-domain and out-of-domain evaluations of neural IR, and seek to improve its robustness across different scenarios, including zero-shot settings. We show that synthetic training examples generated using a sequence-to-sequence generator can be effective towards this goal: in our experiments, pre-training with synthetic examples improves retrieval performance in both in-domain and out-of-domain evaluation on five different test sets.

1 Introduction

Information retrieval (IR) aligns search queries to relevant documents or passages in large document collections. Traditional approaches such as TF-IDF (Salton and McGill, 1986) and BM25 (Robertson and Zaragoza, 2009) rely on simple lexical matching for query-passage alignment. In contrast, neural IR encodes passages and questions into continuous vector representations, enabling deeper semantic matching.

Modern neural IR methods (Lee et al., 2019; Chang et al., 2019) typically employ a dual encoder architecture (Bromley et al., 1993), where separate pre-trained masked language models (MLMs) (Devlin et al., 2019) are fine-tuned for question and passage encoding. Karpukhin et al. (2020) show that effective weak supervision for this fine-tuning step can be derived from existing machine reading comprehension (MRC) datasets (Kwiatkowski et al., 2019; Joshi et al., 2017). Their Dense Passage Retriever (DPR) model demonstrates state-of-the-art (SOTA) retrieval performance on multiple

Wikipedia datasets (Kwiatkowski et al., 2019; Joshi et al., 2017; Berant et al., 2013; Baudiš and Šedivỳ, 2015), with 9–19% performance improvement over term matching baselines like BM25. DPR has also been shown to outperform other neural methods that employ sophisticated IR-specific pre-training schemes such as the Inverse Cloze Task (Lee et al., 2019) or latent learning of the retriever during MLM pre-training (Guu et al., 2020). Recent open domain QA systems (Lewis et al., 2020b; Izacard and Grave, 2020) have successfully used DPR to obtain SOTA performance on end-to-end QA.

However, despite this remarkable success of neural IR in standard supervised learning settings—where the training and test instances are sampled from very similar distributions—it is not clear if and to what extent such an approach would generalize to more practical zero-shot settings. In this paper, we shed light on this question via an empirical study of DPR zero-shot performance on multiple datasets. Our experiments on four out-of-domain datasets show that the advantage over BM25 continues to hold in near domains (datasets with Wikipedia articles comprising the retrieval corpus), but not in a far domain (biomedical text).

Inspired by the success of synthetic training examples in MRC (Shakeri et al., 2020; Zhang et al., 2020; Sultan et al., 2020), we further investigate if a similar approach can be useful for IR, including zero-shot application settings. Concretely, we train a sequence-to-sequence generator using existing MRC data and use it to generate synthetic question-answer pairs from source domain passages. Following the procedure of Karpukhin et al. (2020), we then create IR pre-training examples from these synthetic MRC examples. We observe that pre-training on the generated examples (before fine-tuning with human annotated examples) improves the robustness of DPR: performance consistently improves across all our in-domain and out-of-domain test sets. Importantly, the gap with
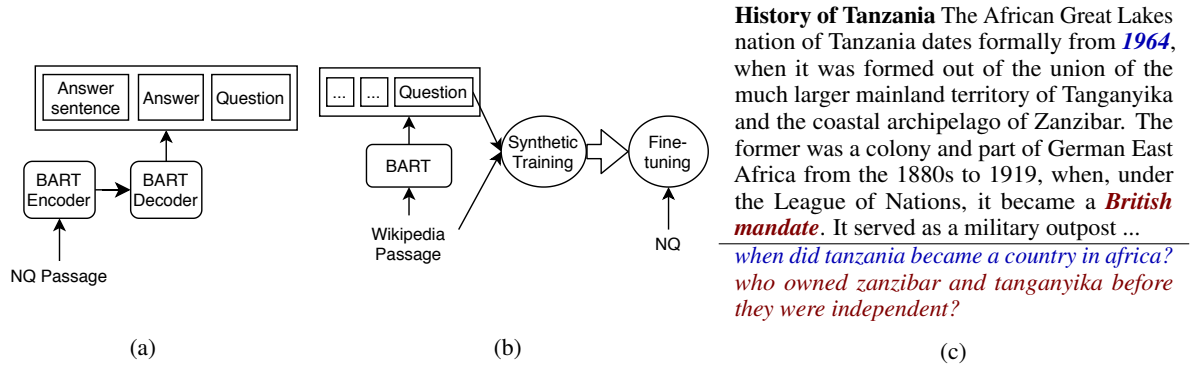
Figure 1: Proposed IR training pipeline along with a synthetic example. (a) A BART encoder-decoder LM is fine-tuned on NQ for QA example generation. (b) Synthetic examples generated from Wikipedia passages are used to pre-train the neural IR model before fine-tuning on NQ. (c) Two synthetic questions output by our generator from the depicted Wikipedia passage, with corresponding answers highlighted in the text.

BM25 in far domain evaluation is also significantly reduced.

The main contributions of this paper are:

- We conduct an empirical evaluation of SOTA neural IR on multiple in-domain and out-of-domain test sets and show how its effectiveness varies across different test conditions.
- We show that synthetic pre-training significantly improves the robustness of neural IR.
- We achieve new SOTA performance in neural IR on five different datasets, including zero-shot settings.

2 Synthetic Pre-Training of IR

Let c be a corpus and $d \in c$ be a document; an IR example consists of a question q and a passage p in d such that p contains an answer a to q. Let s in p be the sentence that contains a.

We train an example generator that selects first a candidate sentence s from an input p, then a candidate answer a in s, and finally generates a corresponding question q. To achieve this, we fine-tune BART (Lewis et al., 2020a), a pre-trained encoder-decoder language model, to generate an ordered triple (s, a, q) from p. Labeled (p, s, a, q) tuples for training are collected from Natural Questions (NQ) (Kwiatkowski et al., 2019), an existing MRC dataset over Wikipedia articles. In practice, we approximate the selection of the answer sentence s by generating its first and last word. Finally, (q, p) is retained as a synthetic IR example. Our method closely resembles that of Shakeri et al. (2020), except that we additionally perform an explicit selection of s.

In our experiments, we generate synthetic training examples from Wikipedia articles. Following (Sultan et al., 2020), we use top-p top-k sampling (Holtzman et al., 2020) to promote diversity in the generated examples. Training and inference for the synthetic example generator are depicted in Figures 1a and 1b respectively. Figure 1c shows two example questions output by the generator from a Wikipedia passage.

For each generated question q, we obtain a negative example by sampling a passage from corresponding BM25 retrievals that does not contain the generated answer a. Following Karpukhin et al. (2020), we use in-batch negatives while training the DPR model. After pre-training with synthetic data, we finally fine-tune the model on IR examples derived from existing MRC data. We name this augmented DPR model *AugDPR*. We refer the reader to (Karpukhin et al., 2020) for a more detailed description of the DPR training process.

3 Experimental Setup

3.1 Datasets

We briefly describe our datasets in this section. Example QA pairs from each are shown in Table 1.

Training and In-Domain Evaluation. We train our systems on Natural Questions (NQ) (Kwiatkowski et al., 2019), a dataset with questions from Google's search log and their human-annotated answers from Wikipedia articles. Lewis et al. (2020c) report that 30% of the NQ test set questions have near-duplicate paraphrases in the training set and 60–70% of the test answers are also present in the training set. For this reason, in addition to the entire NQ test set, we also use the non-overlapping subsets released by Lewis et al.

Dataset	Question	Answers
NQ	what does hp mean in war and order	['hit points or health points']
	who was named african footballer of the year 2014	['Yaya Touré']
TriviaQA	Who was the man behind The Chipmunks?	['David Seville']
	On a standard dartboard, which number lies between 12 and 20?	['five', '5']
WebQuestions	who was richard nixon married to?	['Pat Nixon']
	what highschool did harper lee go to?	['Monroe County High School']
WikiMovies	what does Tobe Hooper appear in?	['Body Bags']
	Mick Davis directed which movies?	['The Match']
BioASQ	Which antiepileptic drug is most strongly associated with spina bifida?	['Valproate']
	Which receptor is inhibited by bimagrumab?	['activin type II receptors']

Table 1: Sample questions from each of the datasets used in evaluation.

(2020c) for in-domain evaluation.

Near Domain Evaluation. For zero-shot near domain evaluation, where Wikipedia articles constitute the retrieval corpora, we use the test sets of three existing datasets.

TriviaQA (Joshi et al., 2017) contains questions collected from trivia and quiz league websites, which are created by Trivia enthusiasts.

WebQuestions (WQ) (Berant et al., 2013) consists of questions obtained using the Google Suggest API, with answers selected from entities in Freebase by AMT workers.

WikiMovies (Miller et al., 2016) contains question-answer pairs on movies, built using the OMDb and MovieLens databases. We use the test split adopted in (Chen et al., 2017).

Far Domain Evaluation. For zero-shot far domain evaluation, we use a biomedical dataset.

BioASQ (Tsatsaronis et al., 2015) is a competition[1] on large-scale biomedical semantic indexing and QA. We evaluate on all factoid question-answer pairs from the training and test sets of task 8B.

Retrieval Corpora. For NQ, TriviaQA, WebQuestions and WikiMovies, we use the 21M Wikipedia passages from Karpukhin et al. (2020) as the retrieval corpus. For BioASQ, we take the abstracts of PubMed articles from task 8A and split into passages of up to 120 words (preserving sentence boundaries). Table 2 shows the sizes of the corpora and the number of questions in each test set.

3.2 Setup

Training. We train the synthetic example generator using the *(question, passage, answer)* triples from NQ. Then we randomly sample 2M passages from our Wikipedia retrieval corpus and generate around 4 synthetic questions per passage. We ex-

Dataset	Domain	Passages	Questions
NQ	Wikipedia	21.0M	3,610
TriviaQA	Wikipedia	21.0M	11,313
WebQuestions	Wikipedia	21.0M	2,032
WikiMovies	Wikipedia	21.0M	9,952
BioASQ	Biomedical	37.4M	1092

Table 2: Statistics of the retrieval corpus and test sets for different datasets used for evaluating the information retrieval models.

perimented with more passages but did not see any improvements on the NQ dev set. Please see appendix for more details on these experiments. For top-p top-k sampling, we use $p = 0.95$ and $k = 10$.

During synthetic pre-training of DPR, for each of the 2M passages, we randomly choose one of its synthetic questions at each epoch to create a synthetic training example. After synthetic pre-training, we fine-tune DPR on NQ to get the AugDPR model. We refer the reader to the appendix for hyperparameter details.

Baselines and Metrics. We use BM25 as our term matching baseline. As a stronger neural baseline, we use the DPR-single model trained on NQ and released[2] by Karpukhin et al. (2020). Both DPR and AugDPR use *BERT-base-uncased* for question and passage encoding. As in (Karpukhin et al., 2020), our evaluation metric is top-k retrieval accuracy, the percentage of questions with at least one answer in the top k retrieved passages.

4 Results and Discussion

Table 3 shows NQ results on the entire test set as well as on the two subsets released by Lewis et al. (2020c). Synthetic pre-training yields larger gains on the non-overlapping splits, with up to 4-point improvement in top-1 retrieval accuracy.

To assess the cross-domain utility of AugDPR,

---

[1]http://bioasq.org/participate/challenges

[2]https://github.com/facebookresearch/DPR

Model	Total			No answer overlap			No question overlap		
	Top-1	Top-10	Top-20	Top-1	Top-10	Top-20	Top-1	Top-10	Top-20
BM25	30.5	54.5	62.5	26.4	47.1	54.7	31.0	52.1	59.8
DPR	46.3	74.9	80.1	32.2	62.2	68.7	37.4	68.5	75.3
AugDPR	**46.8**	**76.0**	**80.8**	**36.0**	**65.0**	**70.8**	**41.4**	**70.8**	**76.6**

Table 3: NQ top-k retrieval results. Performance improves across the board with synthetic pre-training (AugDPR), but more on the non-overlapping subsets of Lewis et al. (2020c).

| Model | Near Domains | | | | | | Far Domain | |
	TriviaQA		WebQuestions		WikiMovies		BioASQ	
	Top-20	Top-100	Top-20	Top-100	Top-20	Top-100	Top-20	Top-100
BM25	66.9	76.7	55.0	71.1	54.0	69.3	**42.1**	50.5
DPR	69.0	78.7	63.0	78.3	69.8	78.1	34.7	46.9
AugDPR	**72.2**	**81.1**	**71.1**	**80.8**	**72.5**	**80.7**	41.4	**52.4**
Supervised	79.4	85.0	73.2	81.4	-	-	-	-

Table 4: Zero-shot neural retrieval accuracy improves with synthetic pre-training (AugDPR) in all out-of-domain test settings. However, BM25 remains a strong baseline on the far domain dataset of BioASQ. Numbers for the supervised models are taken from (Karpukhin et al., 2020).

we evaluate it zero-shot on both near and far domain test sets. Table 4 shows the results. For comparison, we also show numbers for supervised models reported by Karpukhin et al. (2020) on TriviaQA and WebQuestions where the DPR model was trained directly on the training splits of these datasets. For the near domain datasets, we observe that both DPR and AugDPR outperform BM25 by a sizable margin; additionally, AugDPR consistently outperforms DPR. Furthermore, performance of AugDPR on WebQuestions is comparable to the supervised model. On the far domain, however, we observe BM25 to be a very strong baseline, with clearly better scores than DPR. The synthetic pre-training of AugDPR reduces this gap considerably, resulting in a slightly lower top-20 score but a 2-point gain in top-100 score over BM25.

To investigate the relative underperformance of neural IR on BioASQ, we take a closer look at the vocabularies of the two domains of Wikipedia articles and biomedical literature. Following Gururangan et al. (2020), we compute the overlap between the 10k most frequent tokens (excluding stop words) in the two domains, represented by 3M randomly sampled passages from each. We observe a vocabulary overlap of only 17%, which shows that the two domains are considerably different in terminology, explaining in part the performance drop in our neural models. Based on these results, we also believe that performance of neural IR in distant target domains can be significantly improved

via pre-training on synthetic examples that are generated from raw text in the target domain. We plan to explore this idea in future work.

We also examine the lexical overlap between the questions and their passages, since a high overlap would favor term matching methods like BM25. We find that the coverage of the question tokens in the respective gold passages is indeed higher in BioASQ: 72.1%, compared to 58.6% and 63.0% in NQ and TriviaQA, respectively.

Karpukhin et al. (2020) report that DPR fine-tuning takes around a day on eight 32GB GPUs, which is a notable improvement over more computationally intensive pre-training approaches like (Lee et al., 2019; Guu et al., 2020). Our synthetic pre-training takes around two days on four 32GB GPUs, which is comparable with fine-tuning in terms of computational overhead.

5 Conclusions and Future Work

We have shown that pre-training SOTA neural IR models with a large amount of synthetic examples improves robustness to degradation in zero-shot settings. Our experiments show consistent performance gains in five in-domain and out-domain test sets, even in far target domains with significant vocabulary mismatch with the training set. Future work will explore zero-shot domain adaptation of neural IR systems with synthetic examples generated from target domain raw text.

References

Petr Baudiš and Jan Šedivỳ. 2015. Modeling of the question answering task in the yodaqa system. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 222–228. Springer.

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*, pages 1533–1544.

Jane Bromley, James W Bentz, Léon Bottou, Isabelle Guyon, Yann LeCun, Cliff Moore, Eduard Säckinger, and Roopak Shah. 1993. Signature verification using a "siamese" time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, 7(04):669–688.

Wei-Cheng Chang, X Yu Felix, Yin-Wen Chang, Yiming Yang, and Sanjiv Kumar. 2019. Pre-training tasks for embedding-based large-scale retrieval. In *International Conference on Learning Representations*.

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

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.

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*, pages 8342–8360, Online. Association for Computational Linguistics.

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*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.

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)*, pages 1601–1611.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Ledell Wu, Sergey Edunov, Danqi Chen, and Wentau 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, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

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*, pages 6086–6096.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2005.11401*.

Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2020c. Question and answer test-train overlap in open-domain question answering datasets. *arXiv preprint arXiv:2008.02637*.

Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-value memory networks for directly reading documents. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1409.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Gerard Salton and Michael J. McGill. 1986. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., USA.

Siamak Shakeri, Cicero Nogueira dos Santos, Henghui Zhu, Patrick Ng, Feng Nan, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2020. End-to-end synthetic data generation for domain adaptation of question answering systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5445–5460, Online. Association for Computational Linguistics.

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*, pages 5651–5656.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):138.

Rong Zhang, Revanth Gangi Reddy, Md Arafat Sultan, Vittorio Castelli, Anthony Ferritto, Radu Florian, Efsun Sarioglu Kayi, Salim Roukos, Avi Sil, and Todd Ward. 2020. Multi-stage pre-training for low-resource domain adaptation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5461–5468, Online. Association for Computational Linguistics.

Appendix

Hyperparameters

In this section, we share the hyperparameters details for our experiments. Table 5 gives the hyperparameters for training the generator and Table 6 lists the hyperparameters for pre-training and finetuning the neural IR model.

Our BM25 baseline is based on Lucene[3] Implementation. BM25 parameters $b = 0.75$ (document length normalization) and $k_1 = 1.2$ (term frequency scaling) worked best.

Hyperparameter	Value
Learning rate	3e-5
Epochs	3
Batch size	24
Max Sequence length	1024

Table 5: Hyperparameter settings during training the synthetic example generator (BART) using data from NQ.

Hyperparameter	Pre-training	Finetuning
Learning rate	1e-5	1e-5
Epochs	6	20
Batch size	1024	128
Gradient accumulation steps	8	1
Max Sequence length	256	256

Table 6: Hyperparameter settings for the neural IR model during pre-training on synthetic data and finetuning on NQ.

How Many Synthetic Examples do We Need?

To analyze how much synthetic data is required, we experiment with pre-training using 1M and 4M synthetic examples while keeping the number of training updates fixed. As Table 7 shows, we don't see improvements from using more examples beyond 2M.

Model	Top-10	Top-20	Top-100
DPR	73.6	78.1	85.0
AugDPR-1M	74.4	79.2	85.5
AugDPR-2M	74.8	79.7	85.9
AugDPR-4M	74.6	79.1	85.9

Table 7: Retrieval accuracy on the Natural Questions dev set with varying number of synthetic examples (1M vs 2M vs 4M) during pre-training.

---

[3] https://lucene.apache.org