Synthetic Target Domain Supervision For Open Retrieval QA

Revanth Gangi Reddy[1*], **Bhavani Iyer**[2], **Arafat Sultan**[2], **Rong Zhang**[2],
Avirup Sil[2], **Vittorio Castelli**[2], **Radu Florian**[2], **Salim Roukos**[2]

[1] University of Illinois at Urbana Champaign [2] IBM Research AI, New York

*Work done during AI Residency at IBM Research, New York

Open Retrieval Question Answering

Open Retrieval Question Answering (ORQA) involves two steps:

1. **Information Retrieval (IR)**:

   Retrieve relevant passages from a large document collection given the query.

2. **Machine Reading Comprehension (MRC)**:

   Extract the answer spans given the question and the passage.

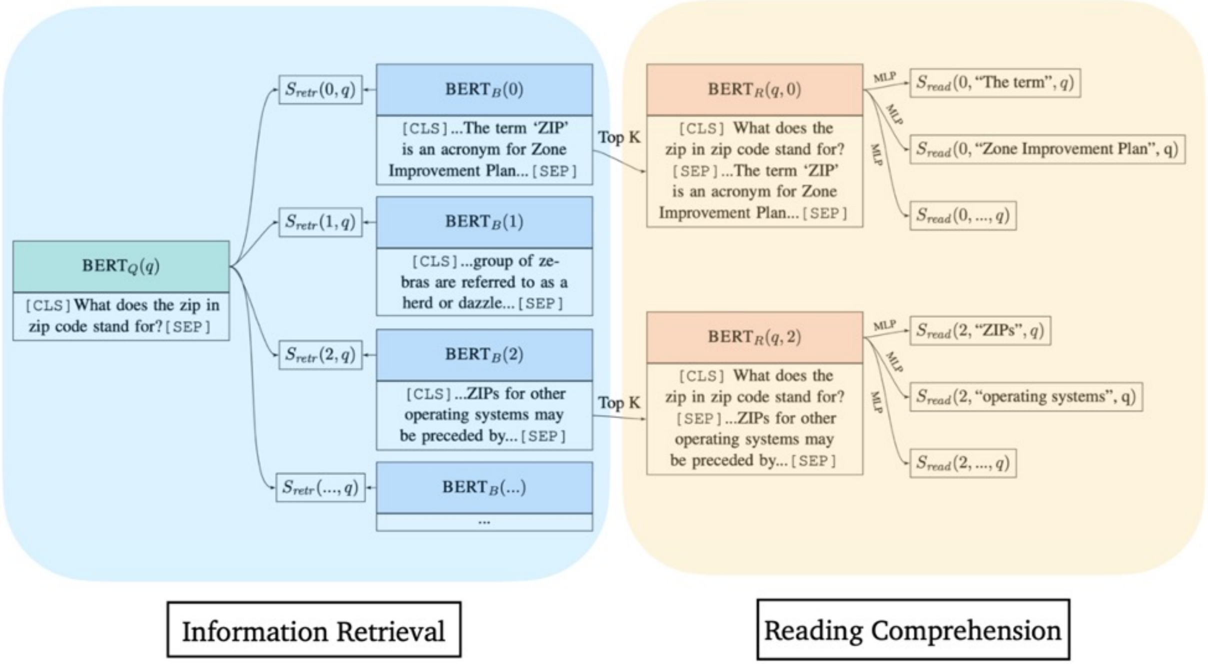


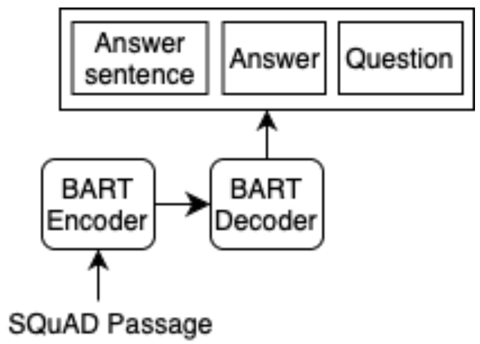
Image taken from *Open Domain Question Answering*, Danqi Chen and Wen-tau Yih, ACL 2020 Tutorial

Contributions

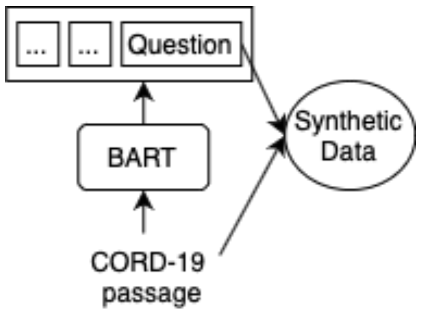
- We empirically show that in out-of-domain Open Retrieval QA (ORQA), advantage of neural IR over BM25 diminishes or disappears.
- We use automatic text-to-text generation to create target domain synthetic training data. Our synthetic examples improve both IR and end-to-end ORQA results.
- Ensembling over BM25 and our improved neural IR model yields the best results.

Synthetic Example Generation

- We finetune a BART[1] model with data from SQuAD[2] to generate synthetic training examples for both IR and MRC.



Generator Training



Generator Inference

- MRC training example is a triple (*passage, question, answer*) and IR training example uses just (*passage, question*).

[1] *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension,* Lewis et al, ACL 2020

[2] *SQuAD: 100,000+ Questions for Machine Reading Comprehension of Text,* Rajpurkar et al, EMNLP 2016

Synthetic Example Generation

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
... As with any research, this study is also not without its limitations. First, is the issue of low response rate despite concerted efforts by the research team to contact key informants multiple times. Scholars have argued that such research is often perceived as opportunistic, by the respondents and this perceived lack of trust is likely to have impacted response rates ...	**Q:** What is the main limitation of this study? **A:** low response rate **Q:** Why was there a low response rate? **A:** perceived lack of trust

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

Datasets

1. COVID-QA-2019[1]:
    - 2019 question-article-answer triples
    - Questions are de-duplicated to create an open version

2. COVID-QA-147[2]:
    - 147 question-article-answer triples with 27 unique questions

3. COVID-QA-111[3]:
    - 111 question-answer pairs

Dataset	IR	MRC	ORQA
COVID-QA-2019	Dev: 201 Test: 1775	Dev: 203 Test: 1816	Dev: 201 Test: 1775
COVID-QA-147	-	Test: 147	-
COVID-QA-111	Test: 111	-	Test: 111

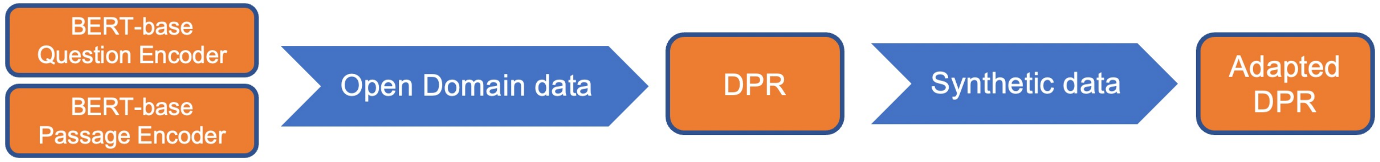
Retrieval Corpus

We use the June 22 version (around 74k documents) of the CORD-19[4] collection. We split the abstract and main body into passages with no more than 120 words.

[1] *COVID-QA: A Question Answering Dataset for COVID-19,* Moller et al 2020
[2] *Rapidly Bootstrapping a Question Answering Dataset for COVID-19,* Tang et al 2020
[3] *Answering Questions on COVID-19 in Real-Time,* Lee at al 2020
[4] *CORD-19: The Covid-19 Open Research Dataset,* Wang et al 2020

Neural IR

- For target domain supervision of the Dense Passage Retriever[1] (DPR), we fine-tune its off-the-shelf open domain instance with target domain synthetic examples.



- *Adapted DPR* considerably outperforms the baseline DPR and BM25 systems on both the datasets.

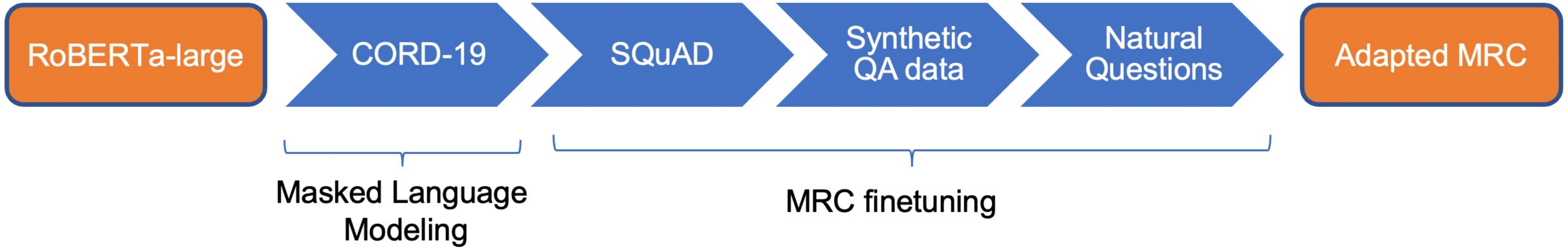
Model	Open-COVID-QA-2019						COVID-QA-111		
	Dev			Test			Test		
	M@20	M@40	M@100	M@20	M@40	M@100	M@20	M@40	M@100
BM25	22.4	24.9	29.9	29.9	33.4	39.7	48.7	60.4	64.9
DPR-Multi	14.4	18.4	22.9	13.8	17.5	21.4	51.4	57.7	66.7
ICT	16.6	21.6	25.5	18.1	23.0	29.6	52.8	59.8	67.6
Adapted DPR	28.0	31.8	39.0	34.8	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**

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

[1] *Dense Passage Retrieval for Open-Domain Question Answering,* Karpukhin et al, EMNLP 2020

Machine Reading Comprehension

- The synthetic QA data are filtered using a roundtrip consistency filter to remove noisy examples and then used in the MRC fine-tuning process.



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**

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

Natural Questions: a Benchmark for Question Answering Research, Kwiatkowski et al, TACL 2019

Open Retrieval QA

- In the open retrieval QA setup, we report numbers from different pairings of IR and MRC systems. We see that both *Adapted DPR* and *Adapted MRC* contribute to improvements in the final F1 scores.

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**

End-to-end F1 scores achieved by different Open Retrieval QA systems.

Zero Shot Evaluation on BioASQ

- Our models also show considerable improvements when evaluated on BioASQ[1] Task 8B.
- These results show that synthetic training on the CORD-19 articles transfers well to the broader related domain of biomedical QA.

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**

IR results on BioASQ Task 8B factoid questions

Model	Top-1	Top-5
BM25 $\longrightarrow$ Baseline MRC	30.6	45.5
DPR-Multi $\longrightarrow$ Baseline MRC	28.6	43.0
Adapted DPR $\longrightarrow$ Baseline MRC	32.1	49.4
Adapted DPR $\longrightarrow$ Adapted MRC	**32.9**	**49.5**

ORQA F1 scores on BioASQ Task 8B factoid questions

[1] *BioASQ: A Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering,* Balikas et al

Conclusion

- We show that synthetically generated target domain examples can support strong domain adaptation of neural open-domain open retrieval QA models.
- Crucially, we assume zero labeled data in the target domain and rely only on open domain machine reading comprehension annotations to train our generator.