Synthetic Target Domain Supervision for Open Retrieval QA

Revanth Gangi Reddy[1], Bhavani Iyer[2], Md 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

Contributions

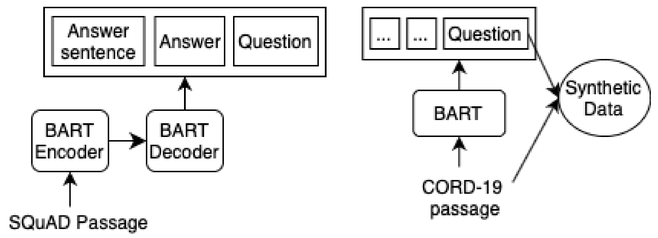
- We empirically show that in out-of-domain Open Retrieval QA (ORQA), advantage of neural IR over BM25 diminishes or disappears altogether in the absence of target domain supervision.
- We use automatic text-to-text generation to create target domain synthetic training data. Our synthetic examples improve both IR and end-to-end ORQA results, in both original and related target domains, requiring no supervision with human annotated examples.
- Ensembling over BM25 and our improved neural IR model yields the best results, which underscores the complementary nature of the two approaches.

Method

Generating Synthetic Examples

To train the generator, we fine-tune BART, a pre-trained denoising sequence-to-sequence generation model, with MRC examples from SQuAD.

Given a target domain passage at inference time, the question-answer pairs are sampled from the generator using top-k top-p sampling.



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 ...	**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

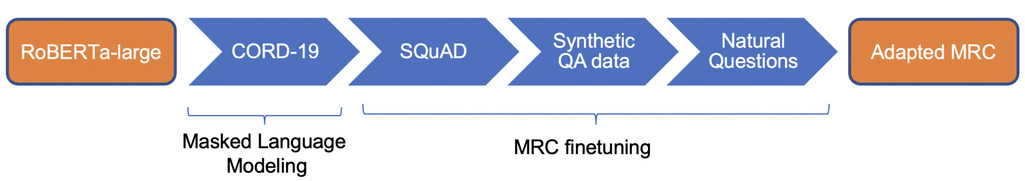
Table: Synthetic MRC examples generated by our generator from a snippet in the CORD-19 collection.

Passage Retrieval

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

Machine Reading Comprehension (MRC)

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



Experiments

Our experiments on passage retrieval show that *Adapted DPR* considerably outperforms the baseline DPR and BM25 systems on both the datasets. Finally, ensembling over BM25 and neural approaches yields the best results with BM25+*Adapted DPR* ensemble being the top system across the board.

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

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**

Table: 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.

Our models also show considerable improvements when evaluated on BioASQ 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**

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

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: ORQA F1 scores on BioASQ 8B factoid questions.

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.