IBM Submissions to EPIC-QA
Open Retrieval Question Answering on COVID-19

Bhavani Iyer[†], Vikas Yadav[†], Martin Franz[†], Revanth Gangi Reddy[‡], Md Arafat Sultan[†], Salim Roukos[†], Vittorio Castelli[†], Radu Florian[†], Avirup Sil[†]

[†]IBM Research AI, T.J. Watson Research Center, New York, USA

[‡]University of Illinois, Urbana Champaign, USA

{bsiyer,franzm,roukos,vittorio,raduf,avi}@us.ibm.com,

{arafat.sultan,vikasy}@ibm.com,

revanth3@illinois.edu

Abstract

This paper describes IBM's submissions to the Epidemic Question Answering (EPIC-QA) challenge at TAC 2020. Given a large document collection on matters related to COVID-19, EPIC-QA asks QA systems to find answers to users' questions in that corpus. We describe our three different submissions, each of which follows a three-stage process of passage retrieval, answer finding and re-ranking. We present implementation details as well as experimental results for all three systems.

1 Introduction

This article describes IBM's submissions to the TAC 2020 Epidemic Question Answering (EPIC-QA)[1] track. EPIC-QA is an open retrieval QA (Karpukhin et al., 2020; Asai et al., 2020) challenge that was designed in response to the COVID-19 pandemic: given a background corpus of text, participating systems were asked to extract answers that are no more than a few sentences long to users' COVID-related questions. In Section 2, we provide a more detailed description of the task.

Solving an open retrieval QA task like EPIC-QA requires both finding documents or passages in a large corpus that are likely to contain an answer to the question, and finding shorter, more specific answer snippets within those passages. Each of our three submissions consists of a separate passage retrieval component (Karpukhin et al., 2020) and an answer finder component (Chakravarti et al., 2020). Additionally, we find that a third re-ranker module that combines the scores of the retriever and the finder can further improve performance. The different implementations of this shared retrieve-find-rerank framework in our three different submissions are detailed in Section 3.

Across our three submissions, we bring together approaches ranging from classical BM25 to cutting-edge NLP techniques like task-specific fine-tuning of large pre-trained transformer neural networks (Liu et al., 2019; Karpukhin et al., 2020). To deal with the highly specialized target domain of COVID-19 and the resulting domain shift for our open domain QA models, we also explore their augmentation with automatically generated target domain training examples (Reddy et al., 2020; Sultan et al., 2020). Details of these approaches are provided in Section 4.

Our experimental results in Section 5 show that a system consisting of i) a passage retriever ensemble of a dense passage retriever adapted to the COVID-19 domain with a neural classifier based retriever, ii) an answer finding component to predict specific answer spans and iii) a final answer sentence re-ranker is consistently the best in both Task A and Task B.

2 Task Description

EPIC-QA consists of two tasks aimed at two different types of users: Task A for subject matter experts and Task B for general consumers. The retrieval corpora consist of: (1) for Task A, the CORD-19 collection of scientific articles (Wang et al., 2020), and (2) for Task B, articles used by the Consumer Health Information Question Answering (CHIQA) service[2] of the U.S. National Library of Medicine (NLM).

For each task, a QA system is to answer users' questions from the respective corpus. An answer generally takes the form of contiguous sentences extracted from a relevant document. Table 1 shows examples of questions and answer sentences from the prelim round of the EPIC-QA challenge. Systems may submit up to 1000 ranked answers per question.

A system is evaluated based on its coverage of manually annotated "nuggets" of key information for the given question: the more such unique

[1] https://bionlp.nlm.nih.gov/epic_qa

[2] https://chiqa.nlm.nih.gov

Question: What is the origin of COVID-19? **Answer:** Recent research shows that ferrets, cats, and golden Syrian hamsters can be experimentally infected with the virus and can spread the infection to other animals of the same species in laboratory. **Nuggets**: { *experimentally infected animals* }
Question: Are there blood tests that detect antibodies to coronavirus? **Answer:** Serology testing is used to detect previous infection (antibodies to MERS-CoV) in people who may have been exposed to the virus. **Nuggets:** { *antibodies*, *antibody test*, *immune response*, *previous infection* }

Table 1: EPIC-QA examples. For each question, we show both an annotated answer and corresponding annotated "nuggets" of key information the answer covers.

nuggets in its top n answers, the higher the score. Table 1 also shows some relevant nuggets for the respective questions. The evaluation metric is a modified form of Normalized Discounted Cumulative Gain (NDCG) called Normalized Discount Novelty Score (NDNS) that considers the number of novel nuggets in an answer sentence. It includes a penalty for sentences that do not contain any of the annotated nuggets and for sentences that only contain already seen nuggets. Further details on the evaluation are provided in Section 5.

3 IBM Submissions to EPIC-QA

For each task, i.e., expert and consumer, we submitted three runs. All three systems adopt a common retrieve-extract-rerank framework, but each implements the retrieval component differently as discussed below. Further details on individual modules are provided in Section 4.

1. **IBM-1:** This system utilizes elastic search (BM25) for passage retrieval, but further reranks the retrieved passages using a neural classifier (Section 4.1). It extracts answer spans from these re-ranked passages using the GAAMA machine reading comprehension (MRC) system (Section 4.4). A final component uses the predicted answer spans across all passages to select and re-rank the answer (Section 4.5).

2. **IBM-2:** The only difference between this system and the IBM-1 system is the passage retrieval module. Here we use an ensemble (Section 4.3) of: (i) the passage retrieval module of the IBM-1 system, and (ii) the neural retriever of Section 4.2. The MRC and answer re-ranking modules remain the same.

3. **IBM-3** This system uses a third passage retrieval algorithm, here we use an ensemble of BM25 and a neural dense passage retriever adapted to the COVID-19 domain (Section 4.2). The MRC and answer re-ranking modules remain the same as the other two systems.

In all three runs, IR returns the top 3000 passages, which we then map to their respective source documents. The top 1000 documents are run through MRC and the final answer re-ranker.

4 System Description

Figure 1 depicts the end-to-end flow as well as the individual components that are combined to create the pipeline for each run. An offline process is run to build the passage level indices used by the the Neural IR components described in Section 4.1 and Section 4.2.

The input question is passed to the IR component to retrieve a ranked list of 3000 passages. For each retrieved passage, we take the full text of the source document as the context for MRC. Each question-document pair is passed to the MRC model to obtain the top 20 short answer spans (e.g., named entities) as described in Section 4.4. These short spans can be thought of as machine equivalents of the annotated "nuggets" of key information for that question. The spans along with their documents are passed to a final combine and re-rank component. Re-ranking is done using a weighted average of the normalized IR score for the source document and the MRC score of the answer span. The final system outputs are the sentences that contain the high-scoring spans.

4.1 Neural Classifier-Based Re-Ranker

We create Elastic Search queries using the "question" field from the task query records, and run them against an Elastic Search index based of passage in the Consumer and Expert document collections. The top 3000 documents from the initial IR round are re-scored using a neural network classifier, similar to the approach described in (Pappas

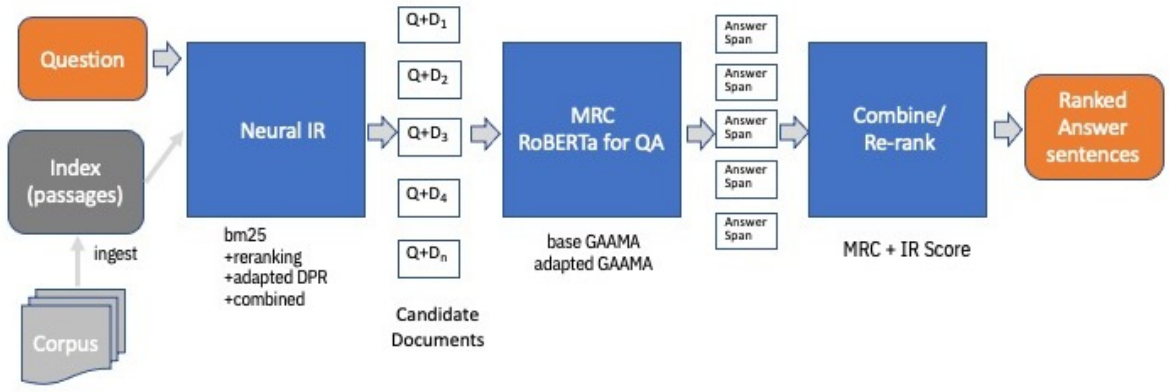

Figure 1: Open Retrieval Question Answering for EPIC-QA.

et al., 2020).

The classifier uses a BERT-based component, output of which in the form of the output vector for the [CLS] token is combined by an additional network layer with features based on the scores and ranks of the retrieved documents obtained from Elastic Search. The network is trained on data and relevance judgements from (1) the BioASQ Task 8b training set,[3] and (2) the preliminary round of the EPIC QA evaluation.

4.2 Adapted DPR

Our neural retrieval model for the IBM-2 system is based on the Dense Passage Retriever (DPR), a state-of-the-art open-domain IR model (Karpukhin et al., 2020) trained on human-annotated MRC data. Following Reddy et al. (2020), we further fine-tune this model on synthetic MRC examples generated from unlabeled text in the CORD-19 collection (Wang et al., 2020). Specifically, we fine-tune BART (Lewis et al., 2020) on examples from SQuAD2.0 (Rajpurkar et al., 2018) and use the scientific articles from CORD-19 to generate synthetic question-passage pairs.

The passages in the EPIC-QA expert and consumer collection are then encoded and stored in an index using FAISS (Johnson et al., 2019). These pre-computed representations are used at inference time to retrieve passages relevant to the question, using nearest neighbour similarity search.

4.3 Ensembling of IR systems

IBM-2 and IBM-3 are both ensemble systems that compute a convex combination of the scores assigned to a passage by two different IR systems as

[3] http://www.bioasq.org/participate/challenges_year_8

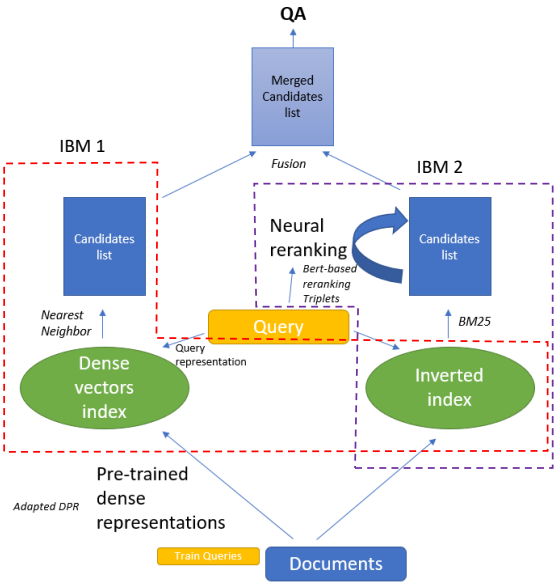

Figure 2: Architecture of our IBM-3 Epic-QA submission. All three submissions share the same QA model (top); the IR components taken from IBM-1 and IBM-2 are shown in red and violet boxes, respectively.

shown in Figure 2. We ran a one-dimensional grid search over $w \in \{0.1, 0.2, 0.3...1.0\}$. The best performing IBM-2 ensemble weighed the BM25 score by 0.4 and thus the DPR score by 0.6. In IBM-3, the best performance was achieved by weighing the neural classifier by 0.7 and the combined BM25+DPR score from IBM-2 by 0.3.

4.4 The GAAMA MRC Reader

Given the collection of retrieved documents, the task of finding more fine-grained answer sentences is performed using a machine reading comprehension (MRC) component. Each document with the question is processed by the MRC model to extract answer spans. The predicted answer span is used

to extract the answer sentence.

We use a RoBERTa (Liu et al., 2019) MRC reader called GAAMA (Chakravarti et al., 2020), which predicts a short answer span in a passage as the answer to a given question. GAAMA fine-tunes a RoBERTa-large language model for QA with an additional task-specific layer (Devlin et al., 2019).

We first train an open-domain instance of GAAMA by fine-tuning on the SQuAD 2.0 (Rajpurkar et al., 2018) dataset for 2 epochs and then on the Natural Questions (NQ) (Kwiatkowski et al., 2019) dataset for 1 epoch (Chakravarti et al., 2020). We also train a second GAAMA instance specifically for the COVID-19 domain by first fine-tuning the RoBERTa language model on the raw text of the CORD-19 document collection and then with synthetic MRC examples generated from the CORD-19 documents, similar to the technique used in AdaptedDPR (Reddy et al., 2020).

For EPIC-QA, we experimented with both GAAMA instances. In our experiments on the EPIC-QA preliminary round question sets, we found the mean NDNS score for the open-domain GAAMA system to be about 1.5 points higher than the adapted GAAMA, therefore we selected the former for submission.

Table 2 shows examples of predicted MRC answer spans and the corresponding predicted answer-bearing sentences for example questions from the EPIC-QA prelim round.

4.5 Answer Sentence Re-Ranker

This component is the final step where answer spans predicted by the MRC component are used to select the answer-bearing sentence and re-rank the sentence for the final submission. We use simple techniques such as ROUGE-1 to measure sentence similarity to remove duplicates. Contiguous sentences are merged and checked to ensure they are within the passage boundaries specified in the document collection. The final score for the selected sentence is the weighted average of the IR score of its document and the MRC score of the highest scoring answer span in the sentence. The weights of 0.7 IR and 0.3 MRC were arrived at via tuning on the COVID-QA dev set (Reddy et al., 2020).

5 Experimental Results

The evaluation results from the preliminary and final rounds are summarized in Table 3. The scores are averages over the 21 questions of Task A and the 18 question of Task B that were judged in the Prelim round, and the 30 questions from the Final round.

Question (EQ001): What is the origin of COVID-19? **Answer Correct:** COVID-19 is a new coronavirus of beta coronavirus genus which originated in bats. **Nuggets**: { *beta coronavirus*, *bats* }
Question (EQ036): What is the protein structure of the SARS-CoV-2 spike? **Answer Correct:** We first collected negative-stain electron microscopic images of sarscov spike protein ectodomain and showed that it is clove-shaped trimer with three individual s1 heads and trimeric s2 stalk fig 1b li et al 2006a **Nuggets:** { *clove-shaped trimer with three individual s1 heads and trimeric s2 stalk* }
Question (EQ005): What drugs have been active against SARS-CoV or SARS-CoV-2 in animal studies? **Answer Incorrect:** Remdesivir has demonstrated in vitro and in vivo activity in animal models against viral pathogens mers and sars which are also coronaviruses and are structurally similar to sarscov2 **Nuggets:** { *Remdesivir* }
Question (CQ033): What vaccine candidates are being tested for COVID-19? **Answer Incorrect:** State and local public health departments have received tests from cdc while medical providers are getting tests developed by commercial manufacturers **Nuggets:** { *commercial manufacturers* }

Table 2: EPIC-QA system generated answers using the IBM 2 system. The first two examples show questions where the top ranked answer is correct. The last two examples show questions where the top ranked answer is incorrect.

The evaluation metric is a modified form of Normalized Discounted Cumulative Gain (NDCG) called the Normalized Discount Novelty Score (NDNS). The scoring includes a Novelty Score (NS) that measures the information in an answer that has not been seen previously in the ranked list.[4]

Formally, the novelty score of an answer is computed as: *NS = (# of novel nuggets) * (# of novel nuggets + 1) / ((# of novel nuggets) + (sentence*

[4] https://bionlp.nlm.nih.gov/epic_qa

Run	Exact	Partial	Relaxed
Prelim Task A			
IBM-1	0.283	**0.330**	0.256
IBM-2	**0.293**	0.259	**0.259**
IBM-3	0.285	0.253	0.254
Prelim Task B			
IBM-1	0.390	0.373	0.372
IBM-2	**0.413**	**0.397**	**0.395**
IBM-3	0.409	0.394	0.393
Final Task A			
IBM-1	0.352	0.330	0.327
IBM-2	**0.367**	**0.345**	**0.344**
IBM-3	0.354	0.336	0.334
Final Task B			
IBM-1	0.266	0.248	0.245
IBM-2	**0.282**	**0.268**	**0.264**
IBM-3	0.278	0.268	0.263

Table 3: Preliminary and Final Round Evaluation

factor)) where a nugget is considered novel if it has not been present in an answer retrieved earlier in the ranked list.

The scorer computes three variants of NDNS:

- Exact: Answers must express a novel nugget in as few sentences as possible and not contain only previously seen nuggets, i.e., *(sentence factor) = (# of sentences in the answer) = (# of sentences with no nuggets) + (# of sentences with seen nuggets) + (# of sentences with novel nuggets).*
- Relaxed: Answers are not penalized for expressing novel nuggets in multiple sentences, but should still not contain sentences with only nuggets provided in previous answers, i.e., *(sentence factor) = (# of sentences with no nuggets) + (# of sentences with seen nuggets) + min((# of sentences with novel nuggets), 1)*
- Partial: Answers are penalized only for sentences with no nuggets, i.e., *(sentence factor) = (# of sentences with no nuggets) + min((# of sentences with novel nuggets), 1)*

Table 3 summarizes the results of the final round of the EPIC-QA challenge. On both Task A and Task B , the IBM-2 system is our best system although the differences between the systems is small. In the Final round, the IBM-2 system is more clearly the best system. On Task A, IBM-2 performed significantly better in the Final round with a mean Exact score of 0.367 versus 0.293 in the preliminary round. Task A, where the retrieval corpus is the CORD-19 collection, clearly benefited from the fine-tuning of the IR component to this domain. Since the consumer collection in the preliminary round was small—around 923 documents—we did not tune the system on that corpus.

6 Conclusion

We have described our submission to the EPIC-QA challenge track of TAC 2020. Our systems follow a three-stage process consisting of passage retrieval, answer finding via a machine reading component and a final answer sentence re-ranker. Given the highly specialized COVID-19 domain where there is not much annotated training data, we explore augmentation with automatically generated examples, and show performance improvements in zero-shot settings.

References

Akari Asai, Jungo Kasai, Jonathan H. Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2020. Xor qa: Cross-lingual open-retrieval question answering.

Rishav Chakravarti, Anthony Ferritto, Bhavani Iyer, Lin Pan, Radu Florian, Salim Roukos, and Avi Sil. 2020. Towards building a robust industry-scale question answering system. In *Proceedings of the 28th International Conference on Computational Linguistics: Industry Track*, pages 90–101, Online. International Committee on Computational Linguistics.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

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*, pages 7871–7880, Online. 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. *CoRR*, abs/1907.11692.

Dimitris Pappas, Ryan McDonald, Georgios-Ioannis Brokos, and Ion Androutsopoulos. 2020. *AUEB at BioASQ 7: Document and Snippet Retrieval*, pages 607–623.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789.

Revanth Gangi Reddy, Bhavani Iyer, Md Arafat Sultan, R. Zhang, Avi Sil, V. Castelli, Radu Florian, and S. Roukos. 2020. End-to-end qa on covid-19: Domain adaptation with synthetic training. *ArXiv*, abs/2012.01414.

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, Online. Association for Computational Linguistics.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Nancy Xin Ru Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The COVID-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.