White Paper

You Want the Truth? Adapt Training Domain to Improve Q&A on Technical Text

Benjamin Roberts, Trent Bradberry,
Will Goodrum, Brittany Pugh

April 2021

ELDER RESEARCH

DATA SCIENCE · AI · MACHINE LEARNING

Table of Contents

1.0 PREAMBLE

Recent advancements in Natural Language Processing (NLP) have been driven by the confluence of large, pre-trained Transformer-based language models (e.g., T5, BERT) and increased availability of and attention to unstructured text data. The International Data Corporation (IDC) predicts that the total volume of existing data will increase over 500% from 2018 to 2025, with 80% of this growth coming from unstructured data.[1] Unlike numerical "big data," text data lacks the necessary structure that typical business analytics datastores such as SQL rely on to efficiently extract valuable insights.

Open-domain Question and Answering (QA) systems provide a solution to this problem, as they allow users to ask questions of their text data in natural language and receive relevant answers without anyone having to specify question and answer pairs in advance. The application of state-of-the-art Transformer models to the task of open-domain QA has resulted in systems that are increasingly accurate and efficient. Further, these tools make large text data accessible to users without technical backgrounds since only natural language queries (rather than programming expertise) are required to extract answers from that data.

Usually, pre-trained language models are built on general language resources, such as Wikipedia articles. While this works well in general knowledge settings (e.g., Internet search), QA performance may degrade on highly technical or domain-specific texts. Additionally, some current QA systems rely on Internet access to generate responses, which may not be viable in Internet-constrained environments (e.g., most defense and intelligence agencies). Flexible and repeatable methods are needed to adapt open-domain QA systems based on large Transformer models to the specialized language of users' text data.

In this whitepaper, we experiment with several recent methods from the NLP literature for adapting state-of-the-art, out-of-the-box, open-domain QA systems to a large, highly technical text corpus. These methods include Domain Adaptive Pretraining and Synthetic QA Fine-Tuning for adapted Machine Reading Comprehension, as well as adapted Dense Passage Retrieval for domain-specific Information Retrieval. We highlight some technical challenges that we encountered in improving performance with domain adaptation, and recommend how to best use these systems in practical settings.

[1] https://www.peakindicators.com/blog/unlocking-insights-from-unstructured-data-with-text-mining

2.0 Problem Statement

2.1 Why Does QA Matter?

Methods that are able to extract valuable information from text data sources are becoming increasingly important. However, identifying key information across thousands of documents covering millions of words can be costly, inefficient, and prohibitive for stakeholders, especially those with non-technical backgrounds. Database tools such as SQL and data wrangling packages like Python Pandas provide methods for querying *structured* tables and data frames. In essence, these tools allow users to first ask questions of the data, and subsequently extract available insights.

QA frameworks offer a similar paradigm to database querying for *unstructured* text corpora. With no technical expertise required, QA tools provide an easy, intuitive framework by which users can enter natural language queries and quickly receive back accurate information from within the text corpus.

2.2 Introduction to Open-Domain QA Using NLP

Standard QA systems extract answers from a text passage that is directly provided to the model. This is the basis of the Stanford Question Answering Dataset (SQuAD)[2]. However, this framework does not help when the answer is located in an unknown document hidden within a corpus of thousands of other documents. Open-domain QA systems address this challenge.

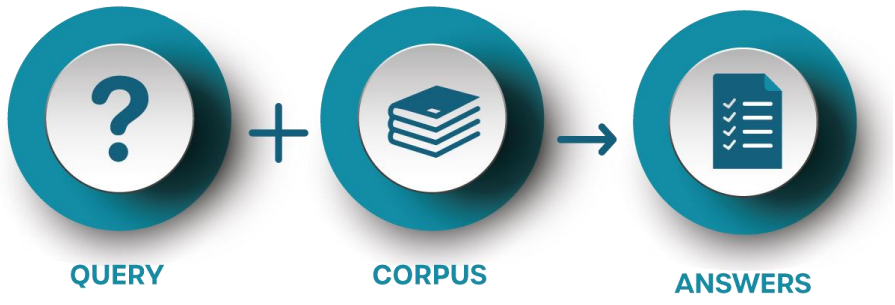

Figure 1. A schematic workflow showing open-domain QA at a high-level

[2] Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. SQuAD: 100,000+ Questions for Machine Comprehension of Text. a*rXiv preprint* arXiv:1606.05250v3, 2016.

Open-domain QA is a specific QA framework that relies on an external knowledge base or text corpus. Open-domain QA involves the interaction between a user-specified query and this external knowledge base, allowing users to extract relevant information and answers available from the knowledge base. State-of-the-art open-domain QA systems typically provide answers in two steps:

1. Find the specific documents that are relevant to the user's query
2. Extract the top answers from these documents

In this way, an open-domain QA system works similarly to the process of conducting research in a library. The answer to a research question can be found on select pages in a handful of books hidden within the stacks of literature the library has to offer. You can find what you are looking for if you look at every book on every shelf, but that's not a great use of time. Instead, it is more practical to pose the research question to the librarian, asking them for direction to the section of the library that contains relevant texts. Once in the right section, you can select and skim or read the books of interest to find relevant information and identify answers to your question (possibly further down-selecting books based on their tables of contents).

In a modern open-domain QA, user-specified queries replace the librarian's expertise on the location of certain information and the reader's efforts to identify answers to their question within relevant texts. By narrowing the required reading material, the open-domain QA framework uses a large corpus of text (knowledge base) to efficiently and accurately provide relevant, valuable information and answers located within it.

Modern open-domain QA systems rely on a two-stage framework first introduced in the DrQA system proposed by Chen et al. (2017)[3] and demonstrated in Figure 2:

1. Information Retrieval (IR)
2. Machine Reading Comprehension (MRC)

[3] Chen, D., Fisch, A., Weston, J., and Bordes, A. Reading Wikipedia to Answer Open-Domain Questions. ar*Xiv preprint* arXiv:1704.00051v2, 2017.

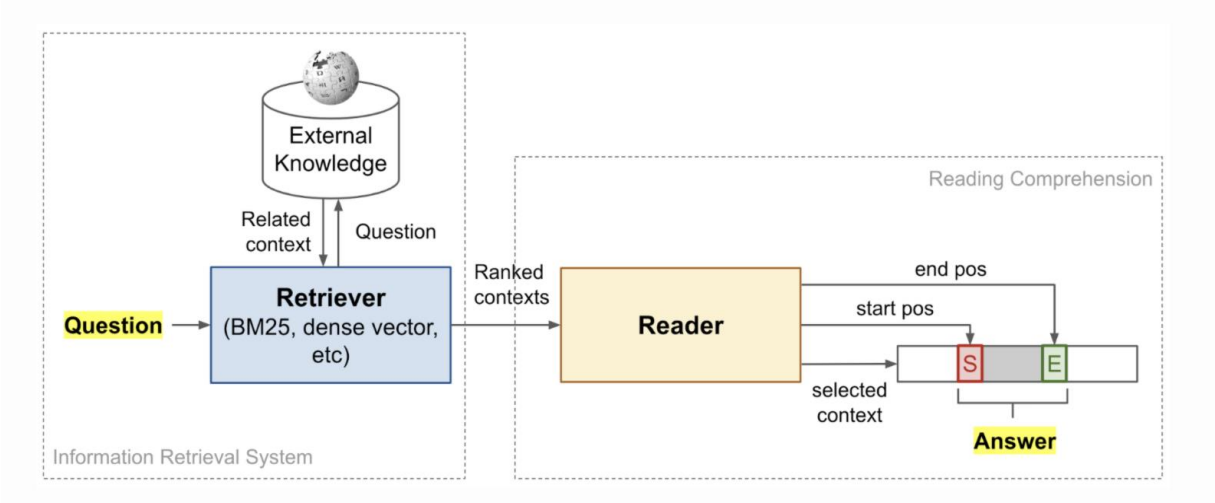

Figure 2. Two-stage framework for open-domain QA[4]

2.2.1 Information Retrieval

The Information Retrieval (IR), or Retriever, stage searches through the corpus and identifies the top-*k* most similar documents relative to a user query. It does so by creating vector representations (or embeddings) of the query and documents, and then comparing them using a distance measure such as cosine similarity. DrQA employed a sparse term frequency - inverse document frequency (TF-IDF) approach to create this embedding, identifying the defining terms that match a given query to a set of documents. Later, Yang et al. (2019) established the BM25 algorithm as the default sparse bag-of-words retrieval approach by demonstrating the algorithm's ability to increase retrieval performance by saturating term frequency and normalizing document length.[5,6]

[4] https://lilianweng.github.io/lil-log/2020/10/29/open-domain-question-answering.html
[5] Yang, W., Xie, Y., Lin, A., Li, X., Tan, L., Tan, L., Xiong, K., Li, M., and Lin, J. End-to-End Open-Domain Question Answering with BERTserini. a*rXiv preprint* arXiv:1902.01718v2, 2019.
[6] https://www.elastic.co/blog/practical-bm25-part-2-the-bm25-algorithm-and-its-variables

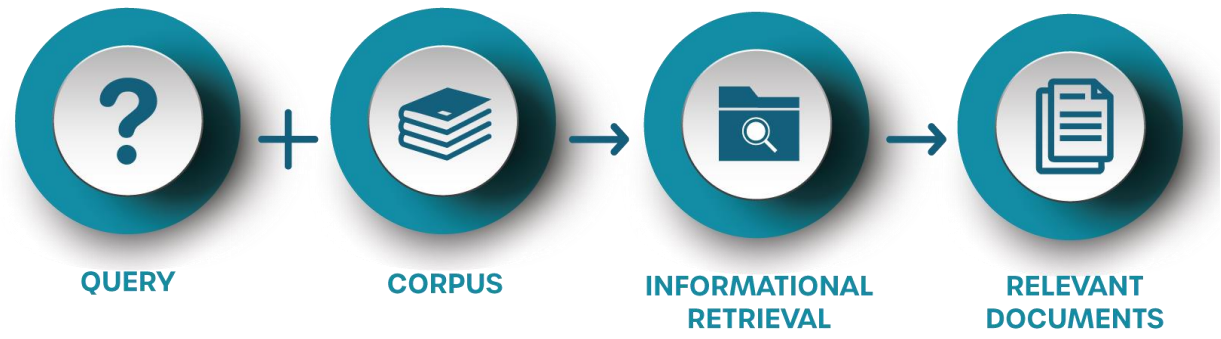

Figure 3. Information Retrieval (IR) component for open-domain QA

2.2.2 Machine Reading Comprehension

The Machine Reading Comprehension (MRC), or Reader, stage then scans through the relevant documents returned by the retriever and returns the top-*n* answers related to the query. Typically, this process relies on a model that is able to identify an answer span by separately predicting the start and end tokens. This is known as extractive MRC, where the answer is pulled verbatim from the text. DrQA used a 3-layer bidirectional LSTM deep neural network to extract these answer spans.

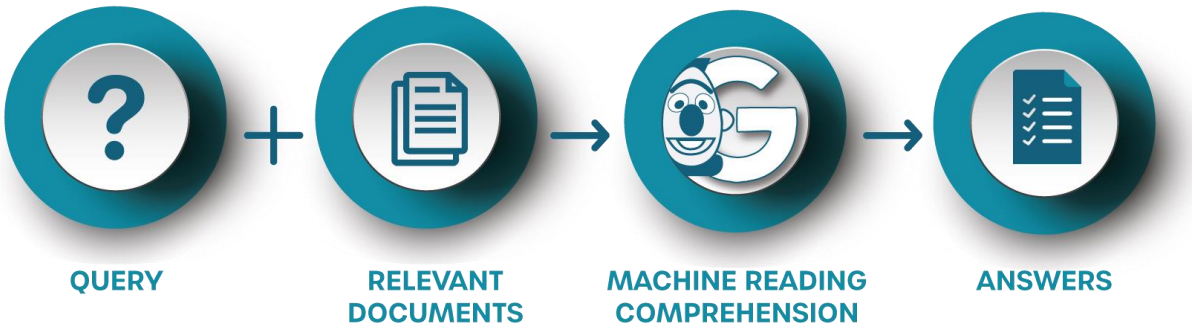

Figure 4. Machine Reading Comprehension (MRC) component for open-domain QA

2.2.3 Advancements in Open-Domain QA

State-of-the-art NLP techniques have enabled drastic improvements recently in the efficiency and accuracy of open-domain QA systems. Specifically, large pre-trained Transformer-based language models such as BERT have improved the ability to search

and read documents through a deep and nuanced “understanding” of context, semantics, and vocabulary.

Transformer-based NLP advancements were first applied to MRC. Spurred by BERT’s dramatic performance increases on the SQuAD benchmark dataset[7], Yang et al. and Wang et al. (2019) successfully applied pre-trained BERT encoders fine-tuned using SQuAD as the Reader model in their performance boosting open-domain QA systems.[8] Concatenating the query and context for input to the model, BERT pays attention to different aspects of the query as it reads through each passage.[9] The BERT Reader framework is demonstrated in Figure 5. More recently, the proposal of Transformer-based encoder-decoder/autoregressive language models such as BART and T5 has allowed for the implementation of generative MRC open-domain QA systems such as RAG and Fusion-in-Decoder.[10,11]

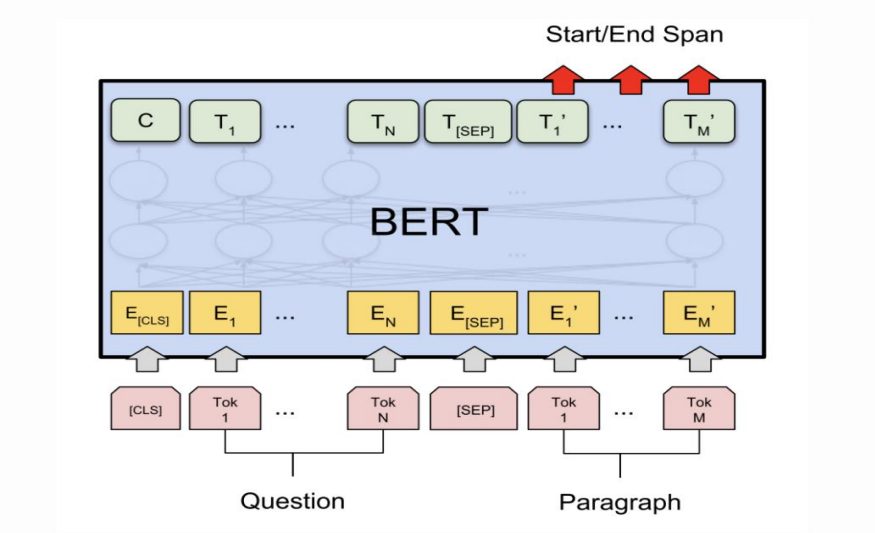

Figure 5. BERT Reader Framework (from [7])

[7] Devlin, J., Chang, M-W., Lee, K., and Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. a*rXiv preprint* arXiv:1810.04805v2, 2019.
[8] Wang, Z., Ng, P., Ma, X., Nallapati, R., and Xiang, B. Multi-passage BERT: A Globally Normalized BERT Model for Open-domain Question Answering. a*rXiv preprint* arXiv:1908.08167v2, 2019.
[9] https://medium.com/deepset-ai/modern-question-answering-systems-explained-4d0913744097
[10] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., and Kiela, D. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. a*rXiv preprint* arXiv:2005.11401v2, 2020.
[11] Izacard, G. and Grave, E. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. a*rXiv preprint* arXiv:2007.01282v2, 2021.

Most recently, BERT-based context embeddings have been applied to the task of Information Retrieval, most successfully through the Dense Passage Retrieval (DPR) framework proposed by Karpukhin et al. (2020).[12] As illustrated in Figure 6, DPR employs a Dual BERT encoder framework to create separate dense embedding representations for the query and context. The framework then identifies the most relevant passages by calculating the cosine similarity between the query and context embeddings. The pre-trained BERT encoders were fine-tuned on an information retrieval task which involved correctly identifying a query's corresponding passage amongst all of the "negative" contexts derived from the other examples in the batch. The resulting embeddings vastly outperformed sparse bag-of-words methods on IR tasks, and the combined DPR-BERT system achieved state-of-the-art performance on open-domain QA benchmarks. DPR retrieval is less efficient than sparse retrieval methods, but this downside is alleviated by calculating all document embeddings prior to querying and using FAISS[13] indexing for efficient similarity search.

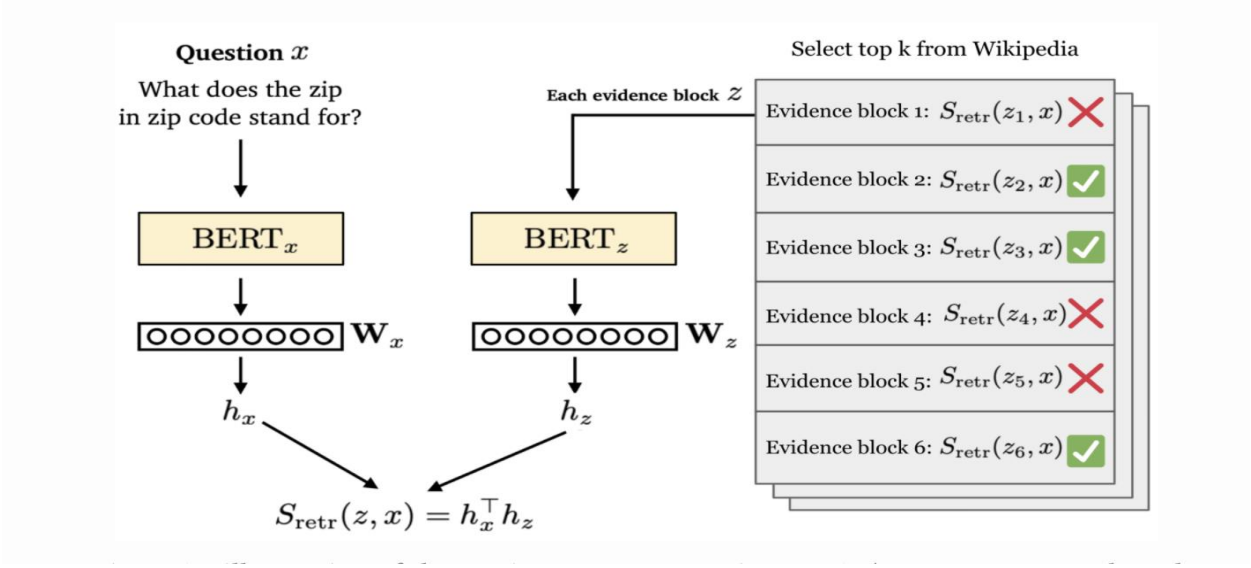

Figure 6. DPR Retriever Framework[14]

[12] Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W. Dense Passage Retrieval for Open-Domain Question Answering. *arXiv preprint*, arXiv:2004.04906v3, 2020.
[13] https://github.com/facebookresearch/faiss
[14] https://github.com/danqi/acl2020-openqa-tutorial/blob/master/slides/part5-dense-retriever-e2e-training.pdf

2.3 Limitations of Current QA Systems

Currently available options for querying large corpora of text face several challenges. Closed-domain QA systems require a relevant text passage to be provided alongside the query for answer extraction -- an impractical requirement for large knowledge bases. Generally speaking, users will not know in advance the questions that they would like answered; questions arise, and then they are asked. Maintaining closed-domain QA systems is also a daunting and resource-intensive task. As organizational needs evolve over time, the relevance of pre-defined QA systems or the ability to specify documents in advance will likely degrade.

While powerful search engines such as Google can make knowledge retrieval from text trivial in many domains, there are contexts where these tools are unavailable (such as in environments where the internet is constrained or denied). Also, search engines will always lack access to proprietary or sensitive data that may be necessary for critical decision making.

Open-domain QA systems can extract answers to user-specified queries, specifically from large collections of proprietary text data. However, such systems that rely on out-of-the-box Transformer-based language models also rely on these models' understanding of language derived from the texts on which they were trained. Since these models are usually trained on generalized language such as Wikipedia articles, their performance tends to degrade on highly technical, domain-specific text. Therefore, methods are required for the reliable and demonstrable domain-adaptation of these highly powerful QA systems for use on technical or domain-specific corpora.

3.0 Methodology

3.1 Proof-of-Concept: Open-Domain QA on Oncological Journals

The goal of our proof-of-concept was to develop a flexible process for adapting a state-of-the-art, out-of-the-box open-domain QA system to a domain-specific corpus. In this way, the benefits of modern NLP advancements in open-domain QA can be applied subsequently to any corpus of text, regardless of the specificity of the domain. Our proof-of-concept had two main technical objectives:

1. Assess the performance of an out-of-the-box open-domain QA system on a large, highly technical corpus
2. Test multiple state-of-the-art domain-adaptation techniques for fine-tuning an open-domain QA system to the same highly technical corpus

For our study, oncology-related research articles from PubMed were collected to compose the text corpus for which the out-of-the-box system would be benchmarked,

and the domain-adaptation techniques would be tested. This corpus was selected due to the accessibility of a large quantity of articles through the PubMed API as well as the highly specific nature of the text. Over 57,000 oncology-related research articles were pulled from PubMed, containing over 240 million words. An example abstract from this corpus is shown in Table 1.

Table 1. Example of Oncology-related PubMed Article from Text Corpus

Title	Abstract
Prediction of the Treatment Response in Ovarian Cancer: A ctDNA Approach[15]	"Ovarian cancer is the eighth most commonly occurring cancer in women. Clinically, the limitation of conventional screening and monitoring approaches inhibits high throughput analysis of the tumor molecular markers toward prediction of treatment response. Recently, analysis of liquid biopsies including circulating tumor DNA (ctDNA) open a new way toward cancer diagnosis and treatment in a personalized manner in various types of solid tumors. In the case of ovarian carcinoma, growing pre-clinical and clinical studies underscored a promising application of ctDNA in diagnosis, prognosis, and prediction of treatment response…"

3.2 Overview of Architecture

The open-source library Haystack[16] is a "framework for building end-to-end question answering systems for large document collections". Specifically, Haystack outlines a two-staged Retriever-Reader open-domain QA system using a user-defined corpus and the latest Transformer-based NLP models (including RoBERTa, ALBERT, etc.), as shown in Figure 7. Using this framework, we established a baseline QA system with an out-of-the-box Retriever and Reader with no adaptation to the oncology domain. To determine the best-performing baseline system, we compared performance between the TF-IDF and DPR retrieval mechanisms. We also tested the effect of varying top-*k* retrieval sizes, comparing the results of 10 and 100 retrieved documents for answer

[15] Sharbatoghli, M., Vafaei, S., Aboulkheyr Es, H., Asadi-Lari, M., Totonchi, M., and Madjd, Z. Prediction of the treatment response in ovarian cancer: a ctDNA approach. *J. Ovarian Res*, v.13; 2020. PMC7574472.
[16] https://haystack.deepset.ai/docs/latest/get_startedmd

extraction. An out-of-the-box Haystack RoBERTa-base model fine-tuned on SQuAD2.0 was used as the baseline reader.[17]

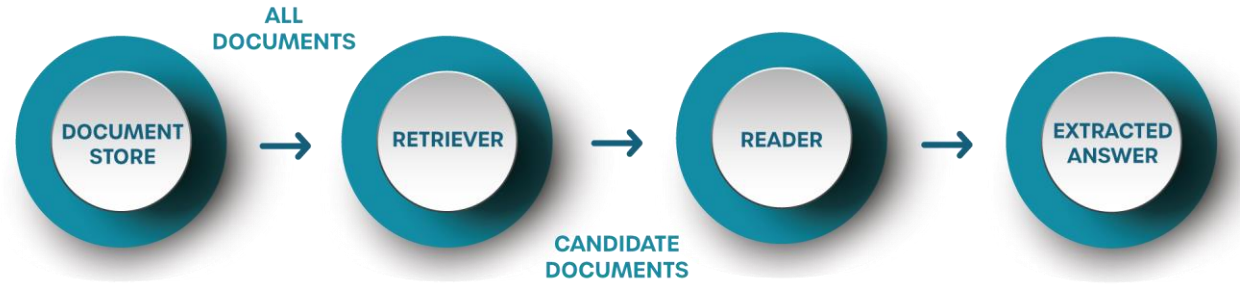

Figure 7. Overview of Haystack Architecture (from [16])

3.2.1 Constructing a QA Validation Dataset

To assess the performance of open-domain QA systems, it was important to develop a validation set of question-answer-text pairs that was both challenging and representative of the highly technical oncology research articles. The SQuAD dataset is the classic QA benchmark, and was created through a time-intensive process of manually annotating question-answer pairs given text entries from Wikipedia. To save time, we leveraged the proposed methods of Alberti et al. (2019) and the Question Generation using Transformers library to produce a validation set of Synthetic QA pairs.[18,19]

Using a T5 model fine-tuned in a multi-task setting, the creation of the Synthetic QA validation set involved the following initial steps:

1. Extract Answer Spans from each passage in the corpus
2. Generate Questions based on these answers and the corresponding passage
3. Perform QA to predict an answer based on the generated question-context pair
4. If the extracted answer span and predicted answer match, add the question-answer-context pair to the Synthetic QA dataset

[17] https://huggingface.co/deepset/roberta-base-squad2
[18] Alberti, C., Andor, D., Pitler, E., Devlin, J., and Collins, M. Synthetic QA Corpora Generation with Roundtrip Consistency. *arXiv preprint* arXiv:1906.05416v1, 2019.
[19] https://github.com/patil-suraj/question_generation#multitask-qa-qg-1

This process was applied to a sample of articles and passages within the PubMed oncology corpus, resulting in over 1 million question-answer-context pairs. By randomly sampling examples from the synthetic dataset and vetting and editing the question-answer pairs to ensure they were both accurate and relatively specific, we created a new validation set containing 74 question-answer-context triplets. This enabled us to create a more robust evaluation set that contained question-answer pairs that were both highly technical as well as specific to our oncology corpus.

3.2.2 Baseline Results

To assess the system's performance on the validation dataset, we used the standard QA literature benchmark metrics Exact Match (EM) and F1 calculated for the top 1 and top 3 answers respectively, as well as two manually annotated metrics. Per the SQuAD paper, the EM score "represents the percentage of predictions that match any one of the ground truth answers exactly", while the F1 score "measures the average overlap between the prediction and the ground truth answer". To allow more flexibility in the predicted answers, we also included a manually annotated Precision@1 score as well as a score indicating whether there was at least one correct answer in the top 3 provided answers. The baseline results are indicated in Table 2. As can be seen, the TF-IDF Retriever with 100 returned documents outperformed both the TF-IDF Retriever with 10 returned documents as well as DPR across all evaluation metrics. The relatively poor performance of DPR indicates the challenges of applying out-of-the-box models to highly technical, domain-specific text. The TF-IDF, *k*=100 Retriever identifies the correct answer in the top 3 extracted results for 69% of the questions in the validation set, compared to 55% for the DPR model. While the sparse TF-IDF Retriever simply relies on a bag-of-words, the DPR Retriever relies on the nuances, semantics, vocabulary, and context of the text. Trained largely on Wikipedia articles, the DPR Retriever faces a steeper learning curve in adapting to the highly technical oncology research articles.

Table 2. Baseline Evaluation Results

	At Least 1 Correct Answer (Top 3)	Precision@1	EM@1	EM@3	F1@1	F1@3
TF-IDF, k=100	**0.69 (51/74)**	**0.57 (42/74)**	**0.31**	**0.18**	**0.41**	**0.29**
TF-IDF, k=10	**0.69 (51/74)**	0.47 (35/74)	0.23	0.13	0.30	0.22
DPR, k=100	0.55 (41/74)	0.41 (30/74)	0.12	0.14	0.20	0.21
DPR, k=10	0.49 (36/74)	0.38 (28/74)	0.11	0.06	0.22	0.16
Google	**0.70 (52/74)**	**0.57 (42/74)**	N/A	N/A	N/A	N/A

In order to acquire a better sense of the absolute performance of our open-domain QA system, we evaluated the performance of Google's search engine on the validation set

across our manually annotated metrics. As indicated in Table 2, Google slightly outperformed our baseline system, correctly identifying the correct answer in the top 3 results for 70% of the validation questions. Google results were identified as correct if Google provided the answer as part of the search engine results, highlighted the answer in the article, or provided a result where the correct answer was easily accessible.

3.3 Experimental Design for QA Domain Adaptation

To improve our baseline and adapt our out-of-the-box, open-domain QA system to the semantics and vocabulary of oncology research articles, we conducted several domain-adaptation experiments involving both the Reader and Retriever. The first two experiments, Domain Adaptive Pretraining (DAPT) and Synthetic QA Fine-Tuning, focused on adapting the Reader model to the nuances of the oncology research article specific text. The final experiment, Adapted Dense Passage Retrieval, involved fine-tuning DPR. The MRC domain-adaptation experiments incorporated the baseline TF-IDF Retriever in the open-domain QA system, while the Adapted DPR experiment utilized the baseline RoBERTa-base Reader fine-tuned on SQuAD2.0.

3.3.1 Machine Reading Comprehension Experiments

Domain Adaptive Pretraining (DAPT) was inspired by the success of Lee et al. (2019), Gururangan et al. (2020), and Reddy et al. (2020) in using a continued pretraining framework to adapt pre-trained language models to tasks based upon domain-specific text.[20,21,22] To conduct DAPT, we divided our PubMed corpus into separate sequences with a maximum length of 512 tokens each. This resulted in over 536,000 separate sequences for training. We then continued pre-training a RoBERTa-base model using the BERT-style Masked Language Model (MLM) objective for 35,000 additional steps. MLM pre-training involves a denoising objective, in which approximately 15% of the words in each sequence are randomly replaced with a [MASK] token, and the model is tasked with predicting the masked words from the surrounding context. In theory, this process enables the RoBERTa Reader to adapt its powerful understanding of language to oncology-research specific terminology and representations. Following the continued pre-training, the domain-adapted RoBERTa-base Reader was fine-tuned on SQuAD2.0 using the same hyperparameters as the Haystack implementation.

Leveraging the methods of Alberti et al. and Reddy et al., the second Reader adaptation experiment involved fine-tuning the model on the 1 million synthetic question-answer-

[20] Lee, J., Yoon, W., Kim, Sung., Kim, D., Kim, Sunk., Ho So, C., and Kang, J. BioBERT: a pre-trained biomedical language representation for biomedical text mining. *Bioinformatics*, 2019, 1-7. doi: 10.1093/bioinformatics/btz682.

[21] Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N.A. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. *arXiv preprint* arXiv:2004.10964v3, 2020.

[22] Reddy, R.G., Iyer, B., Sultan, M.D., Zhang, R., Sil, A., Castelli, V., Florian, R., and Roukos, S. End-to-End QA on COVID-19: Domain Adaptation with Synthetic Training. *arXiv preprint*, arXiv:2012.01414v1, 2020.

context pairs produced from the Synthetic QA generation of the validation set. To ensure the usefulness of this synthetic dataset, we examined a random sample of 100 observations, determining 78% of the examples were valid question-answer pairs based on the context. This enabled us to adapt the open-domain QA system by directly fine-tuning the RoBERTa-base-squad2 Reader model to the task of answering questions related to oncology research. To test the efficacy of this approach, we fine-tuned the model on a random subset of 96,000 examples from the Synthetic QA dataset.

3.3.2 Information Retrieval Experiments

The final round of domain-adaptation experiments focused on adapting the DPR Retriever to the specialized language of the oncology research articles and oncology-specific queries. Since the DPR retriever is a learned method for document encoding, its poor performance in our baseline experiment could be a result of the pre-trained model's unfamiliarity with the highly technical oncology corpus. Thus, based on the methods of Reddy et al., this experiment involved adapting or fine-tuning the DPR query and context encoders to the vocabulary and semantics of the oncology corpus.

Through this experiment, the out-of-the-box Dual BERT Encoder DPR network was fine-tuned using the Synthetic QA dataset developed in earlier experiments. The fine-tuning used in-batch negatives as outlined in Karpukhin et al. and Reddy et al. During training, each query from the synthetic data is accompanied by its corresponding "gold" passage as well as negative passages from the other queries in the batch. Hard negatives, which consist of the top BM25 retrieved passage for each query that does not contain the correct answer, are also incorporated. The fine-tuning process measures negative log likelihood loss of the positive passage to maximize the cosine similarity between the embeddings of corresponding question-context pairs while reducing the similarity between negative pairs. Reddy et al. uses the same hyperparameters from the original experiments to conduct Adapted DPR, with a learning rate of $1x10^{-5}$, batch size of 128, and 6 epochs. However, due to GPU memory constraints, we ran the experiment with a reduced training sample (200k observations), reduced validation sample (25k observations), and reduced batch size (8) for 2 epochs without hard negatives.

4.0 Results and Findings

4.1 Results From Open-Domain QA Experiments

4.1.1 Machine Reading Comprehension Experiments

The results for the Domain Adaptive Pretraining and Synthetic QA Fine-Tuning experiments are displayed in Table 3. The DAPT method resulted in a higher proportion of correct answers identified in the top 3 results with a score of 72%. However, it tied or underperformed both the baseline and Google search results across all other evaluation metrics. Meanwhile, the Synthetic QA Fine-Tuned Reader slightly underperformed the baseline across all metrics except for EM@1.

Table 3. Machine Reading Comprehension Adaptation Results

	At Least 1 Correct Answer (Top 3)	Precision@1	EM@1	EM@3	F1@1	F1@3
Baseline	0.69 (51/74)	**0.57 (42/74)**	**0.31**	**0.18**	**0.41**	**0.29**
DAPT, k=100	**0.72 (53/74)**	0.51 (38/74)	0.26	**0.18**	0.35	**0.29**
Synthetic QA, k=100	0.68 (50/74)	0.45 (33/74)	**0.31**	0.16	0.38	0.22
Google	**0.70 (52/74)**	**0.57 (42/74)**	N/A	N/A	N/A	N/A

The underperformance of the adapted MRC methods could be attributed to multiple factors. In terms of DAPT, the amount of text available based on the size of our corpus was much smaller than that used in previous examples of continued MLM pre-training in the literature, possibly preventing the model from sufficiently adapting to the language of the oncology articles. Hyperparameter choices, especially a low amount of training steps compared with Reddy et al., may also have led to underfitting the model to the oncology corpus. Following continued pre-training with fine-tuning on SQuAD2.0 could have led to catastrophic forgetting of the new semantics and vocabulary acquired during DAPT, especially if there was underfitting due to low amounts of training steps and unique oncology passages.

To address some of these challenges, mixed fine-tuning on both SQuAD2.0 and the Synthetic QA dataset per Reddy et al. might help to alleviate the catastrophic forgetting by providing a domain-related QA task. The DAPT Reader also slightly underperformed Haystack's out-of-the-box RoBERTa Reader on SQuAD2.0 evaluation metrics despite being fine-tuned using the same hyperparameters. This indicates that the increase in domain understanding might have been offset by a decrease in task performance. Finally, building a specialized vocabulary for the Reader prior to continued pre-training and inference might be necessary for highly technical corpora such as this.

In terms of Synthetic QA Fine-Tuning, underperforming may be a result of hyperparameter choices. However, it also might be an issue with the quality of the synthetically generated examples. As mentioned previously, a random sample of synthetically generated question-answer-context triples indicated that 22% of examples also contained invalid question-answer pairs. In these cases, the question did not seem to match the answer based on the context, or the question-answer pair represented a trivial example such as "Which table indicates X fact?". Therefore, it might be necessary to adopt the Roundtrip Consistency approach of Alberti et al. and Reddy et al. to filter

out noisy Synthetic QA examples. This approach uses a pre-trained MRC fine-tuned on SQuAD to determine the highest answerability score over all candidate answer spans for each synthetic question. Synthetic examples are then filtered out based on an answerability threshold tuned using the validation set.

4.1.2 Adapted Dense Passage Retrieval

Following the underperformance of the MRC adaptation experiments compared to the baseline, we turned to adapting the DPR Retriever to the oncology corpus. Given the significant increase in performance DPR provides over a TF-IDF retriever on benchmark QA datasets, we believed fine-tuning DPR on the oncology corpus would provide a significant lift over the baseline system. Ablation studies from Reddy et al. indicate that while an adapted MRC approach through continued pre-training and Synthetic QA Fine-Tuning provides increased performance on domain-specific QA tasks, a larger lift can be attributed to adapted DPR. They indicate a 2.7-7% increase in Top-5 F1 scores across all tasks as a result of adapted DPR, compared to a maximum improvement of 3.7% for adapted MRC. Improved retrieval might be necessary to realize the gains from adapted MRC techniques.

As indicated in Table 4, the Adapted DPR Retriever achieved a significant increase in performance across all evaluation metrics compared to the out-of-the-box DPR Retriever, including an 11% increase in the number of correct answers identified in the top 3 returned results. While the open-domain system including the adapted DPR Retriever still fell short of the baseline, it came within 1-3% across multiple metrics. If we were to fine-tune DPR with the full training sample for the full set of recommended epochs, we believe we might surpass the performance of our baseline system. Further, matching the recommended batch size and including hard negatives in the batch could substantially improve results. Karpuhkin et al. document a 2.3% increase in Top-100 Retrieval accuracy given an increase in batch size from 8 to 128. Adding BM25 hard negatives further increases this performance by 1.8%. Performing Roundtrip Consistency filtering of the Synthetic QA dataset could also improve the DPR fine-tuning process.

Table 4. Adapted Dense Passage Retrieval Results

	At Least 1 Correct Answer (Top 3)	Precision@1	EM@1	EM@3	F1@1	F1@3
Baseline	**0.69 (51/74)**	**0.57 (42/74)**	**0.31**	**0.18**	**0.41**	**0.29**
Adapted DPR, k=100	0.66 (49/74)	0.55 (41/74)	0.23	0.17	0.34	0.27
DPR, k=100	0.55 (41/74)	0.41 (30/74)	0.12	0.14	0.20	0.21
Google	**0.70 (52/74)**	**0.57 (42/74)**	N/A	N/A	N/A	N/A

4.2 Findings and Recommendations for Open-Domain Q&A Implementations

Overall, the adapted MRC improvements did not significantly increase performance over a baseline, out-of-the-box language model. Issues with the quality of the Synthetic QA data, quantity of oncology pretraining data, and hyperparameter choices could all have affected our proof-of-concept results. Nevertheless, research from Reddy et al. indicates that an adapted DPR approach provides a greater lift on domain-specific QA tasks. **IR is the gatekeeper for answer extraction, and without the retrieval of relevant documents, even a domain-adapted MRC will struggle to find a correct answer.**

Our experiments indicate that fine-tuning the DPR Retriever on Synthetic QA data results in a significant increase in performance across all evaluation metrics compared to out-of-the-box DPR. However, due to limitations in the size of the training data, length of training, and batch size, the adapted DPR QA system still falls short of the baseline. Adapted DPR's in-batch negative training process, where the negative examples for each query are drawn from the remaining examples in the batch, requires a large batch size to achieve optimal performance. **Thus, realizing the benefits of the DPR method requires significant GPU memory resources which may be prohibitive for some practitioners**.

For a balance of improved performance and memory practicality **we recommend applying the adapted DPR method with a Roundtrip Consistency filtered Synthetic QA dataset and a reduced batch size**. With the exception of batch size, we recommend following the hyperparameter choices of Karpuhkin et al. and Reddy et al., including 6 training epochs and a learning rate of $1x10^{-5}$. The batch size can be optimized based on the GPU memory resources available.

5.0 Summary

Open-domain Question Answering systems are valuable tools for extracting answers and key insights from unstructured text data. While the application of pre-trained Transformer models to the task of open-domain QA has resulted in dramatic improvements on benchmark datasets, out-of-the-box systems can struggle with highly specialized, domain-specific text corpora. Our baseline system, based on the two-stage TF-IDF Information Retrieval – RoBERTa-base Machine Reading Comprehension framework and implemented using Haystack, identified the correct answer in its top returned 3 results with 69% accuracy. These results were achieved on a validation set of specialized oncology questions synthetically generated from PubMed oncology research articles. Still, Google slightly outperformed the baseline QA system with a score of 70% on the same metric.

Several domain-adaptation techniques from recent literature were tested to adapt the pre-trained language models in the open-domain QA system to the specialized language of the oncology articles and the task of answering questions regarding these articles. Two methods, Domain Adaptive Pretraining (DAPT) and Synthetic QA Fine-Tuning, were tested to adapt the Reader model and improve the system's ability to extract relevant answers from the oncology corpus. Another method, adapted Dense Passage Retrieval, was tested to adapt the Retriever model and improve the system's ability to identify relevant oncology articles for answer extraction. While the MRC and DPR adaptation techniques failed to improve our proof-of-concept system's performance, the adapted DPR technique showed promise given limited GPU resources.

www.elderresearch.com

ELDER RESEARCH

DATA SCIENCE · AI · MACHINE LEARNING