SQuAD to BioASQ: analysis of general to specific

Karen Ouyang
Biomedical Informatics
Stanford University
kjouyang@stanford.edu

Abstract

As biomedical information in the form of publications and electronic health records (EHR) increases at an increasingly fast pace, there is clear utility in having systems that can automatically handle information extraction, summarization, and question answering tasks. While there have been significant strides in improving language tasks for general language, addressing domain-specific contexts still remains challenging. In this project, I apply and fine-tune models to the SQuAD dataset and further modify/adapt for biomedical domain-specific question answering. I evaluated and compared performance on the SQuAD dataset and BioASQ, a biomedical literature QA dataset, with the goal of analyzing and developing approaches to leverage unsupervised language models for domain-specific applications. Upon generating various fine-tuned models, the best performance for general language SQuAD QA achieved an F1 score of 76.717, EM score of 73.379, and for biomedical-specific BioASQ QA achieved an F1 score of 70.348 and EM score of 49.902.

1 Introduction

Machine comprehension tasks have become increasingly important as the amount of content and information being generated is scaling at an unprecedented rate. Question Answering (QA) is a task that has seen accelerated performance improvement with the availability of Stanford's Question Answering dataset (SQuAD) that includes more than 150,000 question-answer pairs and their corresponding context paragraphs from Wikipedia [1]. In just the recent year, performance of QA tasks among many other natural language processing (NLP) tasks encountered a major boost by the advent of pre-trained contextual embeddings (PCE). The objective of PCE is to generate word embeddings that depend on the context in which the word appears in the text opposed to traditional word embeddings such as Word2Vec where each word in the vocabulary is mapped to a fixed vector, regardless of context. PCEs are built by pretraining weights on a large-scale language modeling dataset, then loading the pretrained weights into a model.

While there have been significant strides in improving language tasks for general language, addressing domain-specific contexts still remains challenging. In the scientific biomedical field, the output of publications is estimated to double every 5-10 years, currently with more than 3000 new articles published per day. However, in contrast to general QA, it is challenging to generate a biomedical domain-specific QA dataset comparable to the scale of SQuAD. In this project, I aim to leverage general QA language models for transfer learning of biomedical domain-specific applications.

2 Related Work

The approach to transfer learning and analysis of general QA versus biomedical domain-specific QA in this project is based on modifications and fine-tuning to the language representation model BERT (Bidirectional Encoder Representations from Transformers) [2]. BERT is based on prior

work in pre-training contextual representations. The key is that BERT pre-trains deep bidirectional representations using only a plain text corpus (Wikipedia), by jointly conditioning on both left and right context in all layers. As such, the pre-trained BERT representations can be fine-tuned with just one additional output layer to create models for a wide range of tasks (such as QA) without having to substantially modify model architecture for each specific task. To fine-tune BERT for SQuAD QA, the input question and paragraph were represented as a single packed sequence, with the question and context paragraph using separate embeddings. The only new parameters learned during fine-tuning are a start vector and an end vector, which represent a span of text in the paragraph as the answer. Since the open-sourcing of the BERT model in November, 2018, variations of BERT have continued to top the SQuAD v2.0 leaderboard [3].

3 Approach

The approach for this project is to implement BERT [4] for general language as well as a biomedical domain-specific QA tasks. As illustrated in Figure 1, pre-trained BERT models are fine-tuned with different combinations of general language SQuAD and biomedical domain-specific BioASQ QA [5] training data and hyperparameters. I wrote and added functions to the code base of the different evaluated language models (BERT, BiDAF) to enable training and testing of the BioASQ dataset as well as calculating EM, F1 performance scores. I further compare and analyze the performance of general language (SQuAD) versus biomedical (BioASQ) question answering.

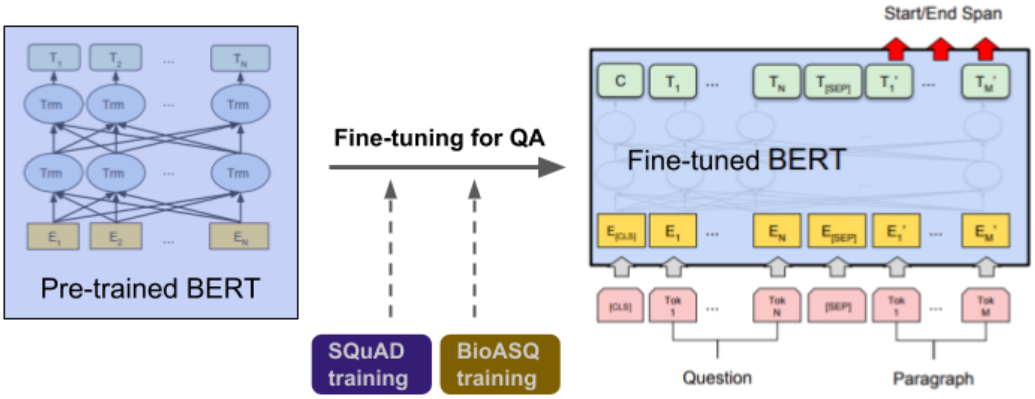

Figure 1: Architecture of BERT for general and biomedical QA

One of the first critical tasks was to evaluate and organize the different data sources into compatible formats. I wrote custom python scripts that took the BioASQ QA data, fetching the appropriate Pubmed publication abstracts (`https://www.ncbi.nlm.nih.gov/pubmed/`) for the context, and modified to be consistent with the SQuAD dataset format.

PCE (BERT)

To compare the performance of general QA with SQuAD versus biomedical-specific QA with BioASQ, I first trained the baseline BERT model with the SQuAD dataset to obtain general baseline performance. I then ran predictions on the BioASQ dataset using the baseline SQuAD model to evaluate how it performs on domain-specific QA. Next, I trained the baseline BERT model with the BioASQ dataset to obtain biomedical baseline performance. As an interesting contrast, I ran predictions on the SQuAD dataset using the biomedical baseline model, to observe how the domain-specific model performed on general QA. With the thought of leveraging the general QA learned from the larger SQuAD dataset, I further trained BioASQ data on a SQuAD-trained model and evaluated performance.

Because BERT was trained on general language corpora (Wikipedia), it likely does not encompass domain specific terms. Thus, I further included the model BioBERT [6], a BERT model additionally pretained on a biomedical-specific corpora containing abstracts from biomedical publications (Pubmed). I then trained BioBERT on SQuAD for general QA and BioASQ for biomedical-specific QA, and assessed performance as described above.

Non-PCE (BiDAF)

I also wanted to evaluate and compare the performance of biomedical QA with a non-PCE model. I trained the baseline BiDAF (Bidirectional Attention Flow for Machine Comprehension) model [7,8] with the SQuAD dataset to obtain general baseline performance. I then ran predictions on the BioASQ dataset to determine how it performs on biomedical-specific QA. Further, I trained a BiDAF model on the BioASQ data to evaluate biomedical-specific performance. Additionally, I generated a BiDAF model that utilizes a biomedical word2vec in place of Glove word vectors to train and test the BioASQ dataset.

4 Experiments

4.1 Data

BioASQ

Biomedical domain-specific question answering dataset was obtained from `http://bioasq.org/`. The BioASQ dataset consists of several different types of language task challenges. For specific question answering task, I wrote custom python scripts to collect questions, biomedical publication abstract context, and answer snippets. Interestingly, in contrast to SQuAD where a single paragraph context contains several different questions, the BioASQ dataset has multiple different publication abstract contexts pertaining to one question. While there were approximately 2000 questions in the BioASQ dataset, I compiled more than 20,000 question-answer pairs (train set: 24,559; dev set: 6,140) indicating that each question has about 10 corresponding publication abstract contexts. Furthermore, the context paragraphs and answer spans for the BioASQ datasets are significantly lengthier than those for SQuAD.

Example:

"question": "Which miRNAs could be used as potential biomarkers for epithelial ovarian cancer?"
"context": "OBJECTIVE: MicroRNA (miRNA) is an abundant class of small noncoding RNAs that act as gene regulators. Recent studies have suggested that miRNA deregulation is associated with the initiation and progression of human cancer. However, information about cancer-related miRNA is mostly limited to tissue miRNA. The aim of this study was to find specific profiles of serum-derived miRNAs of ovarian cancer based on a comparative study using a miRNA microarray of serum, tissue, and ascites.: From 2 ovarian cancer patients and a healthy control, total RNA was isolated from their serum, tissue, and ascites, respectively, and analyzed by a microarray. Under the comparative study of each miRNA microarray, we sorted out several miRNAs showing a consistent regulation tendency throughout all 3 specimens and the greatest range of alteration in serum as potential biomarkers. The availability of biomarkers was confirmed by qRT-PCR of 18 patients and 12 controls.: Out of 2222 kinds of total miRNAs that were identified in the microarray analysis, 95 miRNAs were down-regulated and 88 miRNAs were up-regulated, in the serum, tissue, and ascites of cancer patients. Among the miRNAs that showed a consistent regulation tendency through all specimens and showed more than a 2-fold difference in serum, 5 miRNAs (miR-132, miR-26a, let-7b, miR-145, and miR-143) were determined as the 5 most markedly down-regulated miRNAs in the serum from ovarian cancer patients with respect to those of controls. Four miRNAs (miR-132, miR-26a, let-7b, and miR-145) out of 5 selected miRNAs were significantly underexpressed in the serum of ovarian cancer patients in qRT-PCR.: Serum miR-132, miR-26a, let-7b, and miR-145 could be considered as potential candidates as novel biomarkers in serous ovarian cancer. Also, serum miRNAs is a promising and useful tool for discriminating between controls and patients with serous ovarian cancer."
"answer": "Serum miR-132, miR-26a, let-7b, and miR-145 could be considered as potential candidates as novel biomarkers in serous ovarian cancer"

SQuAD:

The SQuAD dataset consists of more than 150,000 question-answer pairs and the corresponding context paragraph. The SQuAD dataset for this project is split into three parts, train (129,941 examples), dev (6,078 examples), and test (5,915 examples) sets. The questions and answers for the train and dev set are openly available, whereas answers to the test set are entirely secret.

4.2 Evaluation Method

The primary evaluation methods for QA on SQuAD and BioASQ datasets are calculating:

- Exact Match (EM) score
- F1 score (measures the weighted average of precision and recall at the token level)

4.3 Experimental Details

BERT:

To determine and compare the performance of SQuAD general language QA versus BioASQ biomedical-specific QA, I first generated the following fine-tuned BERT models for QA tasks (trained full BERT language model (all layers), learning rate=5e-5, epochs=3, batch-size=12, null-score-diff-threshold=0.0, gradient-accumulation-steps=0.0):

- BERT-S (BERT trained on SQuAD training data)
- BERT-B (BERT trained on BioASQ training data)
- BERT-SB (BERT trained on SQuAD then BioASQ)
- BioBERT-S (BioBERT trained on SQuAD)
- BioBERT-B (BioBERT trained on BioASQ)

I further ran and evaluated the performance of each model on the dev sets for SQuAD and BioASQ, respectively (see Table 2 below for results).

Based on the results of the experiments performed on the different models described above, additional fine-tune adjustments were made to further improve performance of the BioASQ QA task. In analysis of BioASQ output, it was noticed that predicted answers were uniformly shorter than the ground truth answers (Figure 2).

- **Question:** Is there an association between Klinefelter syndrome and breast cancer?
- **Context:** There appear to be no substantial data to confirm the assumption that breast cancer in men with Klinefelter's syndrome is as common as breast cancer in the normal female population. The number of reported cases of breast cancer in Klinefelter's males is only 27, a number too small for any meaningful statistical analysis There is evidence, however, to suggest that Klinefelter's males have an increased risk of breast cancer that approaches three percent. Physicians should therefore be aware of potential breast pathology in XXY males and incorporate a careful breast examination and specific education into the routine health maintenance of men with Klinefelter's syndrome.
- **Answer:** There is evidence, however, to suggest that Klinefelter's males have an increased risk of breast cancer that approaches three percent.
- **Prediction:** Klinefelter's males have an increased risk of breast cancer that approaches three percent.

Figure 2: Tensorboard Text panel showing predicted answer is shorter than ground truth

Additionally, a high proportion of "correct" answers were clustered in the upper half of the context paragraphs compared to the latter half. To address these findings, I adjusted the max-answer-length and max-sequence-length parameters for the BioASQ QA tasks, creating two additional models:

- BERT-SB max-ans=200
- BERT-SB max-ans=200 max-seq=500

BiDAF:

The BiDAF model was trained with the SQuAD dataset using default parameters to set the baseline. Following, I trained the BiDAF model using the BioASQ dataset to set a baseline for biomedical-specific QA. Additionally, I experimented with using a biomedical-specific word vector in place of GLoVE.

4.4 Results and Analysis

The performance of the BiDAF model in predicting SQuAD and BioASQ QA serves as a non-PCE baseline and comparison point of non-PCE versus PCE models for question-answering tasks in general (Table 1). The baseline BiDAF model performs decently, but as expected, has substantially lower scores than the PCE models to be described below.

Additionally, I compared the performance of using a biomedical domain-specific word vector in place of GLoVE, a general language word vector. There was significant improvement in performance for the BioASQ QA task (F1 scores increasing from 33.99 to 40.78, and EM scores improving from 5.71 to 8.70), indicating that using domain-specific word vectors is worthwhile for domain-specific QA tasks.

Table 1: Comparison of SQuAD and BioASQ performance on BiDAF models

Model	Training data	Test data	DevNLL	F1	EM	AvNA
BiDAF-SQuAD	129,941	6,078	3.03	61.64	58.31	68.64
BiDAF*-SQuAD	129,941	6,078	3.23	57.38	53.25	66.07
BiDAF-BioASQ	24,559	6,140	5.32	33.99	5.71	n/a
BiDAF*-BioASQ	24,559	6,140	4.43	40.78	8.70	n/a

* indicates model used biomedical word2vec in place of GLoVE.

The results of running each model on the dev sets for SQuAD and BioASQ (Figure 3), respectively, demonstrate that the fine-tuning process makes a significant impact on the performance of the QA task. General and biomedical-specific QA tasks perform better when fine-tuned with their respective training data. Additionally, it is worth noting that BioBERT (pretrained on general language *and* biomedical domain corpora) fine-tuned on BioASQ training actually performed substantially better than BERT fine-tuned on BioASQ training for SQuAD general language QA (increasing more than 20 points for F1 score). This suggests that a language model, such as BERT, pretrained with a diverse corpora may be able to increase the average performance of general and various domain-specific language tasks.

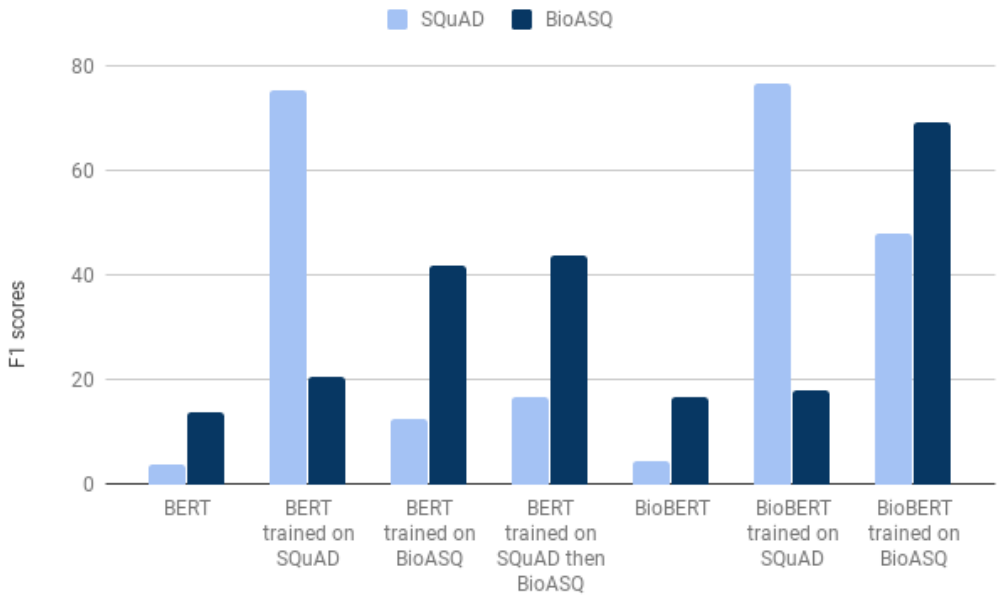

Figure 3: Comparison of SQuAD and BioASQ performance on fine-tuned BERT models

In reviewing the performance of BioASQ QA (Figure 3), I noticed that the performance was not much better than the non-PCE BiDAF model. This prompted additional analyses of the results whereby it was shown that predicted answers for BioASQ QA were on average significantly shorter than the ground truth answers. Furthermore, the "correct" answers tended to be in the front half of the context paragraph compared to the latter half. These findings prompted further fine-tuning of sequence lengths and answer lengths (Figure 4). Comparing `[BERT-SB]` and `[BERT-SB, max-ans=200, max-seq=500]` models, there is a boost of performance, increasing the F1 and E1 scores by more than 20 points, respectively. With improved performance due to adjusting the sequence length and answer length hyperparameters, the performance of BERT trained on BioASQ is now comparable to BioBERT trained on BioASQ.

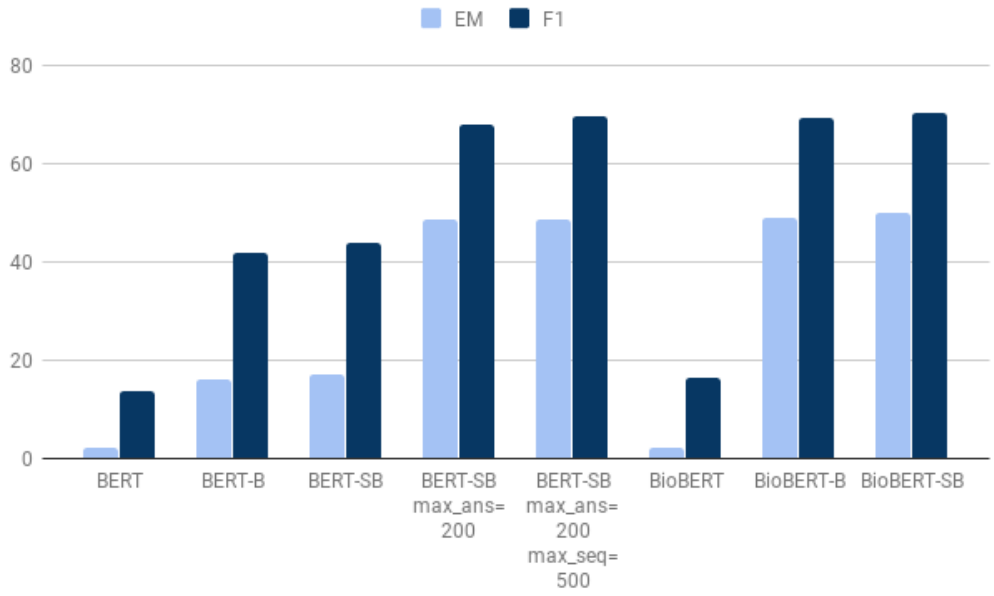

Figure 4: Comparison of Models Predicting BioASQ QA

The best performing fine-tuned models for SQuAD and BioASQ (Table 2), respectively, were models that were trained with their respective training datasets. While BioASQ is significantly lagging SQuAD in exact match answers, the F1 scores are much closer together. Given that BioASQ has approximately 20 percent of the training data of SQuAD, but has a less than 10 percent difference in F1 score, demonstrates that leveraging BERT fine-tuned for domain-specific QA tasks is a great stride.

Table 2: Best performing fine-tuned BERT models for SQuAD and BioASQ QA

Model name	Description	EM	F1
BERT-S	SQuAD dev set	71.997	75.349
BioBERT-S	SQuAD dev set	73.379	76.717
BERT-SB, max-ans=200, max-seq=500	BioASQ dev set	48.615	69.503
BioBERT-SB, max-ans=200, max-seq=500	BioASQ dev set	49.902	70.348

5 Conclusions and Future Work

In this report, I generate multiple fine-tuned BERT models to assess and compare performance of general language SQuAD QA and biomedical-specific BioASQ QA. While the results show that the best performing models for SQuAD and BioASQ are those fine-tuned with their respective training datasets, it is encouraging that with only twenty percent of BioASQ training data compared

to SQuAD, the best-performing fine-tuned model was able to achieve an F1 score that was within 10 percent difference from SQuAD's best-performing F1 score. These results demonstrate that leveraging an unsupervised language model, BERT, for domain-specific QA with substantially less supervised training data achieves results that are nearing comparable to general language QA.

Synthesizing and summarizing information across multiple sources is a core task that biomedical researchers and scientists perform daily. For future work, I aim to take the different predicted "answers" for a question that queries multiple paragraph contexts, and output a summary of the information as well as flag when there is ambiguity or contradiction in the returned answers.

6 Acknowledgements

Thanks to the CS224N course staff for all their help and Microsoft Azure for providing compute resources.

7 References

[1] P Rajpurkar, R Jia, and P Liang. Know what you don't know: Unanswerable questions for squad. arXiv preprint arXiv:1806.03822, 2018.

[2] J Devlin, MW Chang, K Lee, and K Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[3] SQuAD leaderboard. https://rajpurkar.github.io/SQuAD-explorer/ (accessed 3/18/2019).

[4] Pytorch pretrained BERT. https://github.com/huggingface/pytorch-pretrained-BERT

[5] G Tsatsaronis et al. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. BMC Bioinformatics, 16, 138, 2015.

[6] J Lee, W Yoon, S Kim, D Kim, S Kim, CH So, J Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. arXiv preprint arXiv:1901.08746

[7] M Seo, A Kembhavi, A Farhadi, and H Hajishirzi. Bidirectional attention flow for machine comprehension. arXiv preprint arXiv:1611.01603, 2016.

[8] CS 224N Default Final Project: Question Answering on SQuAD 2.0.
https://github.com/chrischute/squad