Reading Comprehension on Covid-19 Research Dataset

Experimental Protocol

Pablo Marino
Catherine Wang
Vishay Vanjani

1 Hypothesis

We hypothesize that to improve BERT's[3] performance on scientific machine comprehension, also known as question answering tasks, the model should be pre-trained on scientific biomedical texts **and** use a domain specific vocabulary. Current state of the art model (BioBert[4]) uses the general vocabulary, same as original BERT, and it did not show significant improvement both on SQuAD2.0 or BioASQ challenge.

The other medical domain model SciBert[2] has shown that having a medical vocabulary helps improve the performance on certain downstream scientific NLP tasks NER, REL, e.g..However, the authors did not evaluate its performance on scientific QA tasks.

Our goal is to solve tasks in the "Kaggle covid-19 open research challenge (CORD-19)"[6] using SciBert fine-tuned with in-domain datasets like BioASQ and out of domain datasets like SQuAD[5]. We also plan to pre-train SciBert on a subset of the CORD-19[6] dataset.

2 Data

There are two candidate QA datasets for fine-tuning our model, BioASQ[1] and SQuAD[5]. In addition, we will use a subset of the CORD-19 research papers in the pre-training process to further enhance the performance.

BioASQ QA is a domain specific dataset available at http://participants-area.bioasq.org/datasets/. Every year, BioASQ releases a new QA dataset as part of the BioASQ challenge[1]. We plan to use BioASQ8 which has 3,243 questions in the training set and will have approximately 700 questions in the test set. The dataset contains 4 types of questions namely Yes/No, Factoid, List & Summary. Each question is annotated with multiple pubmed abstracts(context) and answer texts with start and end offsets.We will process the dataset to 1) Split the context whenever it has more than 512 tokens 2) Create multiple {question,context,answer} tuples for each question since each question in training set has multiple context and answers.

The SQuAD dataset consists of more than 150,000 question-answer pairs with context paragraphs. One third of the questions in the dataset are unanswerable. One paragraph in SQuAD has several question, answer pairs.

3 Metrics

The team will be using the universally accepted machine comprehension and QA evaluation metrics to quantify the model performance and will provide comparison with the original BERT & BioBERT. That evaluation will include the following two scores:

- Exact Match (EM) score, which presents the number of answers that are precisely correct (with the same start and end index of the answer span)
- F1 score, which captures the harmonic mean of precision and recall, of the sequence predicted by the model w.r.t the ground truth answer

For the evaluation on the kaggle COVID-19 task, the team will self evaluate the extracted answers using a manual review process.

4 Models

We use the SciBERT model that is pre-trained on 1.14M papers from Semantic Scholar using original BERT[3] code. 82% of these papers are from the Biomedical domain and 18% are from the computer science domain. The model uses SciVocab, a new WordPiece vocabulary built using SentencePiece library. The token overlap between original Bert's vocabulary and SciVocab is 42% which illustrates a substantial difference in the frequently used words. We will use the uncased model for the QA task.

5 General Reasoning

Recent advances in language modeling have led to substantial gains in the field of natural language processing, with state-of-the-art models such as BERT, RoBERTa, XLNet, and Albert, among many others.

The pre-trained models inherit the characteristics from transfer learning and leverage them on solving problems in the downstream NLP tasks. Those models have been trained with a large amount of unlabeled general-purpose corpus to build the best-distributed representation of word vectors, then fine-tuned on specific NLP tasks.

SciBERT is one step ahead of BERT since it has been pre-trained on scientific papers to build up the vocabulary and scientific embeddings. This gives a considerable improvement on a few NLP tasks in the biomedical domain (including NER, PICO, REL, etc.).

As far as we know SciBERT has never been fine-tuned on a biomedical domain QA dataset like BioASQ. We believe SciBert will perform better than BioBert because it uses a domain specific vocabulary. To further improve SciBert, we will be feeding more biomedical documents(from CORD-19 dataset) to pre-train the model, try different masking techniques and experiment with different model initialization techniques.

6 Summary of Progress

Using the library hugging face we set up two baseline QA models: BERT fine tuned on SQUAD V2, SCIBERT and used them to predict answers in contexts taken from covid-19 articles to perform a qualitative analysis. Tried to evaluate the models on entire datasets like squad 2 to obtain quantitative metrics but had some issues when using the hugging face prediction/evaluation script and are working on that now
Regarding the datasets already downloaded BioAsq V8 and created an account in their page which will allow us to download older versions of their dataset. Found a data mining project[7] in the Covid-19 Kaggle competition[6] which will allow us to create an indexed database out of its huge dataset and plan to use it to find relevant articles to create the SCIBERT pre-training dataset.
Next steps are, obtain metrics for the baseline models, fine tune them on BioAsq, generate a Covid-19 dataset and use it to pretrain SCIBERT, obtain metrics for new models and compare also perform qualitative analysis in Covid-19 QA comparing results across all models.

7 References

[1] George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artiéres, Axel Ngonga, Norman Heino, Eric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos and Georgios Paliouras. 2015. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. BMC bioinformatics.
[2] I Beltagy, K Lo, A Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. arXiv:1903.10676
[3] J Devlin, MW Chang, K Lee, and K Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805

[4] J Lee, W Yoon, S Kim, D Kim, S Kim, CH So, J Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. arXiv:1901.08746
[5] P Rajpurkar, R Jia, and P Liang. 2018. Know what you don't know: Unanswerable questions for squad. arXiv:1806.03822
[6] https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge
[7] https://www.kaggle.com/dirktheeng/anserini-bert-squad-for-semantic-corpus-search