BioMedBERT: A Pre-trained Biomedical Language Model for QA and IR

Souradip Chakraborty[*]
Walmart Labs
souradip24@gmail.com

Ekaba Bisong[*]
SiliconBlast
ekaba.bisong@siliconblast.com

Shweta Bhatt[*]
Jupiter
shwetabhatt08@gmail.com

Thomas O. Wagner III
TPHS Research
thomas.wagner@bertbiomed.org

Francesco Mosconi
Zero to Deep Learning
f@mosconi.me

Riley D. Elliott
TPHS Research
riley.elliott@bertbiomed.org

Abstract

The SARS-CoV-2 (COVID-19) pandemic spotlighted the importance of moving quickly with biomedical research. However, as the number of biomedical research papers continue to increase, the task of finding relevant articles to answer pressing questions has become significant. In this work, we propose a textual data mining tool that supports literature search to accelerate the work of researchers in the biomedical domain. We achieve this by building **BioMedBERT**, a neural-based deep contextual understanding model for Question-Answering (QA) and Information Retrieval tasks. We also leverage the new BREATHE dataset which is one of the largest available datasets of biomedical research literature, containing abstracts and full-text articles from ten different biomedical literature sources on which we pre-train our BioMedBERT model. Our work achieves state-of-the-art results on the QA fine-tuning task on BioASQ 5b, 6b and 7b datasets. In addition, we observe superior relevant results when BioMedBERT embeddings are used with Elasticsearch for the Information Retrieval task on the intelligently formulated BioASQ dataset. We believe our diverse dataset and our unique model architecture are what led us to achieve the state-of-the-art results for QA and IR tasks.

1 Introduction

The COVID-19 pandemic reminded us of the need for a tool that biomedical researchers can use to sift through existing research to extract novel insights, and ultimately help them make novel drug discoveries. The rate of new publications in the biomedical field is on the rise. PubMed reports that more than 1 million biomedical research papers are published each year, amounting to nearly two papers per minute (Landhuis, 2016). For papers mentioning COVID-19 alone, as of June 2020 more than 8000 peer-reviewed publications have been published on PubMed. With the rate of scientific papers on COVID-19 doubling every fourteen days (Coren, 2020), it is imperative to have a language understanding tool that can extract relevant information from credible literature, such as the research methodology, data, authors, results, and citations (Hao, 2020).

In this paper, we address the problem from an information retrieval perspective, extracting the textual and contextual information from the corpus by taking a hierarchical approach. Traditional search approaches such as Lucene-based Elasticsearch (Gormley and Tong, 2015) using BM-25 & Jaccard-based matrices are efficient in retrieving objective answers where the primary task is to extract specific parts of

[*] Denotes first author with equal contribution.
Github repo: https://github.com/BioMedBERT/biomedbert.

Proceedings of the 28th International Conference on Computational Linguistics, pages 669–679
Barcelona, Spain (Online), December 8-13, 2020

the passage. However such methods struggle in the contextual retrieval of documents for which we need latent space representation of the query and the corpus of passages.

Our work leverages the BERT language model architecture to pre-train a large-scale biomedical language representation model, named BioMedBERT. The work is inspired by the research of (Lee et al., 2019) from Korea University & Clova AI research group. In said work, we use the new BREATHE dataset which combines full text articles and abstracts from ten data sources in the biomedical domain and use it to train our BioMedBERT model. We use the $BERT_{LARGE}$ model as a pre-training backbone to achieve new state-of-the-art results for question answering in the biomedical domain. In addition, we obtained impressive results by combining BioMedBERT embeddings with Elasticsearch to obtain highly relevant results for information retrieval. This is achieved by using a neural passage re-ranking mechanism, which learns the inherent structural dependencies in the query and the research articles. We validated our search algorithm by formulating BioASQ as a retrieval dataset.

2 Related Work

Latent space representation learning and vector space modeling have proven to be extremely successful in the natural language processing domain, where they have shown to efficiently encapsulate the hidden meaning and context of sentences or passages. The journey of distributed word representation learning began with the efficient Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014) and fastText (Joulin et al., 2016) which outperformed the traditional bag of words model by significant margins in linguistic tasks, as these methods "largely" ignored the word context of the tokens. With the advent of sequential modeling and recurrent neural models, there was a significant enhancement in information content of the latent representation learning, with LSTMs as the state of the art at the time (Greff et al., 2016). Non-recurrent sequence-to-sequence encoder-decoder models, also known as Transformers, were introduced by (Vaswani et al., 2017). Transformers stacked attention layers into a multi-headed attention architecture. The attention mechanism makes it possible to learn long-running word dependencies between input and output sequences by computing a context vector of the encoder for each token in the sequence (Bahdanau et al., 2014). The BERT architecture, which is primarily the encoder part of the transformer, achieved notable model performances for several linguistic tasks. Pre-training and fine-tuning the BERT network on a domain specific corpus has also shown to outperform many language models (Devlin et al., 2018). Our work combines ideas from (Devlin et al., 2018) and (Lee et al., 2019) in using deep bidirectional transformers to learn contextual embeddings of word features from the large BREATHE corpora.

3 BREATHE Dataset

Biomedical Research Extensive Archive To Help Everyone, or BREATHE, is a new dataset collection of biomedical research articles from leading medical archives. It is a combination of both full body texts and abstracts. The development and recent availability of this dataset significantly inspired the work in this paper. To the best of our knowledge, BREATHE is the largest ***diverse*** collection of publicly available, machine readable biomedical corpus for advanced language modeling (Goncharov et al.,).

The dataset collection process was done in line with "ethical" principles for scraping, along with public APIs that were used when available. The primary advantage of the BREATHE dataset for our model is its source diversity. BREATHE contains full text articles and abstracts from nine sources including BMJ, arXiv, medRxiv, bioRxiv, CORD-19, Springer Nature, NCBI, JAMA, and BioASQ (Goncharov et al.,).

We performed our experiments with BREATHE v1.0 dataset. BREATHE v1.0 contains more than 6M articles and about 4 billion words. BREATHE v2.0 is the most recent version (Goncharov et al.,) and the results reported in this paper are from training BioMedBERT on BREATHE v1.0 dataset. Table 1 summarizes the BREATHE dataset.

Corpus	BREATHE v1.0		BREATHE v2.0	
	Num. of articles	Num. of words	Num. of articles	Num. of words
BMJ	19	41K	47	103K
arXiv	29	5.7K	71	14K
medRxiv	738	443K	1,831	1.1M
Nature Research	12,873	437M	32,376	1.1B
bioRxiv	18,603	2.7M	44,272	6.4M
CORD-19	27,753	9.2M	67,285	22.4M
Springer Nature	322,374	69M	808,265	173.2M
NCBI	828,096	3.5B	2,080,168	9.3B
BioASQ	4,920,278	2.5M	12,123,125	6.4M
Total	**6,130,763**	**4.1B**	**15,090,155**	**10.6B**

Table 1: Corpora details for pre-training BioMedBERT.

4 Methodology

BioMedBERT was built on the foundation of the BERT architecture (Devlin et al., 2018). In our work, we leverage both the pre-training and fine-tuning aspects of the BERT architecture which enhances accuracy while also ensuring the robustness of the model. BERT leverages a transformer architecture (Vaswani et al., 2017) and uses the stacked encoder where each encoder consists of a multi-headed attention layer and a feed-forward network.

BERT builds on the transformer architecture to train large unlabeled data over two self-supervised tasks, and they are the Masked Language Model (MLM) and Next Sentence Prediction (NSP). MLM (also referred to as the *Cloze* task) (Taylor, 1953) works by randomly masking 15% of the sequence and then attempting to predict the masked tokens, where as the NSP task helps the model to understand the relationship between two sentences. It works by predicting if a sentence is the actual "following" sentence or if it is just a random sentence. For more details on the training procedure for BERT, the reader is referred to (Devlin et al., 2018). In training the BERT model, the total loss is computed as the sum of the masked-LM loss and next sentence prediction loss, both of which are cross-entropy, with the former multi-class log-loss and the later binary log-loss.

$$L = \underbrace{\left(-\frac{1}{N}\sum_{i=1}^{N} y.log(p) + (1-y)log(1-p)\right)}_{\text{NSP loss}} + \underbrace{\left(-\sum_{i=1}^{N} q(y).log(q(y))\right)}_{\text{MaskedLM loss}}$$

BioMedBERT Model Architecture. Our model architecture retrained the $BERT_{LARGE}$ architecture with 24 transformer blocks, with a hidden-size of 1024 and 16 attention heads.

4.1 Pre-training BioMedBERT

The "de facto" BERT model was pre-trained on BooksCorpus (800M words) (Zhu et al., 2015) and English Wikipedia (2.5B words) (Devlin et al., 2018). The BioBERT model from (Lee et al., 2019) was pre-trained on different combinations of text-data from English Wikipedia, BookCorpus, PubMed Abstracts (4.5 billion words) and PubMed Central full-text abstracts (13.5 billion words). We trained the BioMedBERT model on BREATHE containing over 6 million articles from 9 different archives with 4 billion words. In pre-processing our dataset, we treated lists, tables and headers as a contiguous sequence of text and initialized our model using the pre-trained BERT weights. Initially we trained the model from scratch with a custom SentencePiece vocabulary for tokenization, but this did not yield good results when evaluated on downstream fine-tuned tasks. We represented the input sequences as vectors using the WordPiece embeddings, which contains 30,000 tokens (Wu et al., 2016). Using WordPiece allowed us to better leverage the pre-trained weights of BERT. WordPiece has a clever formulation to account for

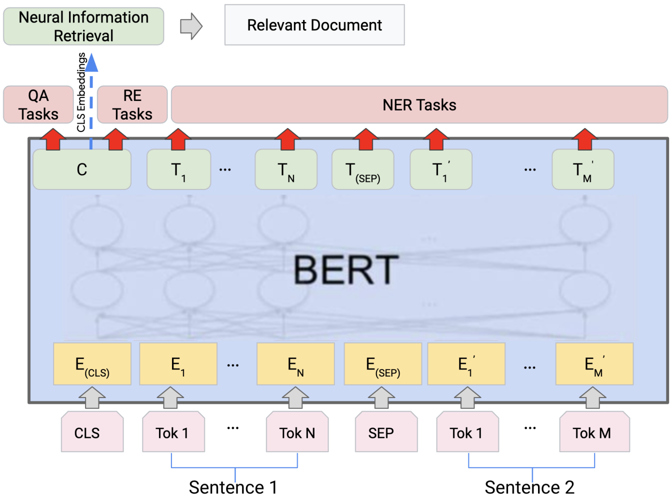

Figure 1: BioMedBERT Information Retrieval Architecture

words not found in its vocabulary by breaking a word into subwords, thereby creating multiple tokens for a given word. Also, WordPiece has a subword for every character in the alphabet.

4.2 Fine-tuning BioMedBERT

The BioMedBERT model retains the architectural effectiveness of the BERT model, which itself leverages the multi-head self attention mechanism of the transformer network which makes it amenable to task specific fine-tuning. It is important to note the [CLS] token assigned by the WordPiece tokenizer, which is a special classification token to represent the aggregate sequence for downstream classification tasks. In building on the framework provided by (Lee et al., 2019) and (Devlin et al., 2018), our model is fine-tuned on three NLP tasks: Named Entity Recognition (NER), Relation Extraction (RE) and Question Answering (QA).

Named entity recognition is the most fundamental sub-task of information extraction, it involves classifying and recognizing named entities such as drugs, diseases, etc. from unstructured biomedical text. The single output layer predicts the token level BIO2 probabilities for each input sequence. We use precision, recall, and F1 score as metrics for evaluation. The details of the datasets used for NER fine-tuning task are mentioned in Table 2.

Dataset	Entity type	Number of annotations
NCBI Disease (Doğan et al., 2014)	Disease	6881
BC5CDR (Li et al., 2016)	Disease	12694
BC5CDR (Li et al., 2016)	Drug/Chem.	15411
BC4CHEMD (Krallinger et al., 2015)	Drug/Chem.	79842
BC2GM (Smith et al., 2008)	Gene/Protein	20703
JNLPBA (Kim et al., 2004)	Gene/Protein	35460

Table 2: Statistical overview of biomedical NER datasets.

Relation extraction is another sub-task of information extraction which involves detection and classification of relationships between named entity mentions in a biomedical corpus. In processing the results for the relation extraction datasets, we incorporated the technique from BioBERT (Lee et al., 2019) in using predefined tags to anonymize target named entities. For evaluation of RE, we also used precision, recall, and F1 score as metrics. Statistical details of the biomedical RE datasets are provided in Table 3.

Question answering is the task of predicting the text span of the answer, given the question and passage containing the answer. For general language question answering task, we fine-tuned BioMedBERT on SQuAD v1.0 and v2.0 datasets using the same BERT architecture used for SQuAD. For the biomedical QA fine-tuning task, we used BioASQ factoid datasets, as their data format is similar to that of the

Dataset	Entity type	Number of relations
GAD (Bravo et al., 2015)	Gene-Disease	5330
EU-ADR (Van Mulligen et al., 2012)	Gene-Disease	355

Table 3: Statistical overview of biomedical RE datasets.

SQuAD datasets (Lee et al., 2019). For evaluating BioASQ, we used the evaluation code from BioBERT to exclude samples with unanswerable questions from the training set. And like (Lee et al., 2019) and (Wiese et al., 2017), we fine-tuned BioASQ on weights that were initially fine-tuned on SQuAD v1.1. However, unlike them, we also fine-tuned BioASQ on SQuAD v2.0 weights which was beneficial in outperforming state-of-the-art results for these tasks. **This is another key contribution in this work**. The results are reported in Table 6 and the dataset details are provided in Table 4.

Dataset	Number of questions
BioASQ 4b-factoid (Tsatsaronis et al., 2015)	1307
BioASQ 5b-factoid (Tsatsaronis et al., 2015)	1799
BioASQ 6b-factoid (Tsatsaronis et al., 2015)	2251
BioASQ 7b-factoid (Tsatsaronis et al., 2015)	2747

Table 4: Statistical overview of biomedical QA datasets.

5 Neural Information Retrieval & Biomedical Search

In this phase, the primary objective was to retrieve the most relevant research papers and articles in a ranked order based on the query made by the researcher(q_i). There are two major aspects that enhance the complexity of this problem. First, the computational complexity associated with retrieval from a biomedical corpus of millions of records(N). Second, the ability to accurately retrieve the relevant documents in a proper rank-ordered fashion, and to do so the process needs to account for a deep contextual understanding of the text.

We approach the problems in a 2-step hierarchical fashion where in the first step, we primarily reduce the search space complexity by using the Elasticsearch and BM-25 algorithm, a well-known metric in retrieval scenarios (Gormley and Tong, 2015). Elasticsearch itself operates in two stages. The first stage creates an extremely efficient data structure in textual searching scenarios, an inverted index, of the entire corpus of biomedical research papers. The second stage queries the entire corpus using the inverted index and scores the search results using the similarity (relevancy) ranking function BM25, and returns the top k search results based on that $BM25$ score, where k is a tunable parameter (Turney and Pantel, 2010). The output of this stage is an ordered state of retrieved biomedical papers sorted by $BM25$ scores and let $(a_{r1}, a_{r2}, a_{r3} ... a_{rk})$ be the top k retrieved documents.

$$BM25(t,d) = \frac{(k+1) \times (t,d)}{(k \times (1 - b + b(\frac{|d|}{|d|_{avg}})))} \times idf(t,D)$$

Here, $f(t,d)$ represents the raw frequency of the term t in the document d, $idf(t,D)$ is a function of number of documents, the term t has occurred in, given the universal set of documents D , $|d|$ is the number of words in the document and k, b are parameters.

The results of the Elasticsearch based retrieval mechanism did not incorporate the contextual aspects of the query and the specific biomedical aspects of the corpus. As mentioned earlier, for effective IR in this domain, contextual consideration is highly relevant. When biomedical researchers search for relevant studies, they search for specific combinations of topics and scenarios. For instance, they might search for “how does coronavirus impact the lungs and to what extent” rather than a generic search such as “what is coronavirus”. In order to best answer these types of specific queries, the system needs robust context for both the question and the possible answers.

To solve this challenge, in the second step of the hierarchical search we use BioMedBERT to project the query(q_i), retrieve the top k papers($a_{r1}, a_{r2}, a_{r3}...a_{rk}$) in a d dimensional latent space and extract the embeddings (Nogueira and Cho, 2019). Let the query embedding vector of q_i be represented by $Q_i \in R^d$ and the matrix embedding representation of the top k retrieved papers be $A_{k\times d}, A_i \in R^d$. We compute the cosine similarity as a vector-matrix product between the normalized query ($Q'_i = \frac{Q_i}{||Q_i||}$) and the normalised retrieved paper embedding matrix ($A'_{k\times d}, whereA'_i = \frac{A_i}{||A_i||}$). Let Z be the output after computing the cosine similarity between the query and the papers, where $Z = A'_{k\times d} \cdot Q'_i$ and $Z \in R^k$. Finally, we sorted the output cosine scores Z , while re-ranking and returning the most relevant biomedical research papers conditioned on the query(q_i).

$$\underset{j\in k}{\operatorname{argmax}} Z_j$$

The primary intent of selecting cosine similarity as the preferred metric is that the vector cosine scores are normalized on each of the dimensions and hence are robust to scaling (Eibl and Gaedke, 2017).

6 Results

6.1 Datasets

In order to simplify the process of comparing our work with related works, we perform the fine-tuning experiments on the biomedical and general language datasets that are most widely used by other NLP researchers. For NER fine-tuning task, we used the eight pre-processed datasets provided by (Wang et al., 2019) mentioned in Table 7. The NER evaluations for all datasets mentioned in Table 7 are based on entity level exact matches. For the RE fine-tuning task, we used the pre-processed GAD and EU-ADR datasets from (Lee et al., 2020) which contain gene-disease relations. For general domain QA fine-tuning task, we used SQuAD v1.1 and SQuAD v2.0 datasets (Rajpurkar et al., 2016) and for biomedical domain QA task, we used pre-processed 4b, 5b and 6b BioASQ datasets provided by (Lee et al., 2020) and pre-processed BioASQ 7b dataset from (Yoon et al., 2019). We report the results in the form of micro-average scores of all five test batches of BioASQ datasets for QA fine-tuning and 10-fold cross validation scores for GAD and EU-ADR dataset for RE fine-tuning task.

6.2 Experimental Setup

In pre-training BioMedBERT, we use the setup provided by BERT (Devlin et al., 2018). For fine-tuning and evaluation, we use the setup provided by (Lee et al., 2020). During the initial phase of experimentation, we explore training the BioMedBERT model from scratch on BREATHE using our custom vocabulary created with SentencePiece tokenizer (Kudo and Richardson, 2018). But after evaluation on a downstream tasks of NER and QA, we found that we need to either add general domain data to our dataset or use the weights of BERT to initialize the model.

Hence, we trained our model using $BERT_{LARGE}$ weights as a transfer learning backbone on the BREATHE dataset. We use the same architecture and hyper-parameters as BERT and trained the model for 68k epochs. The resulting improvements in results was very close to state-of-the-art scores for NER. This motivated us to continue work on BioMedBERT and eventually train the model for 1M steps. The pre-training of BioMedBERT model on Google Cloud v3-TPUs with 128 cores for 1M epochs took a little over 3 days. Fine-tuning tasks on the same TPU took less than an hour.

6.3 Experimental Results

We provide the results for the selected downstream tasks in Tables 5, 6, 7 and 8. In each table, we compare BioMedBERT's performance against BERT and the extant state-of-the-art model for the corresponding task. Among the four fine-tuning tasks selected for evaluation and comparison with previous works, BioMedBERT achieves better results than the BERT model for nearly all of them on biomedical datasets. BioMedBERT outperforms the state-of-the-art models on QA fine-tuning task using SQuAD v2.0 dataset and also achieves close to state-of-the-art results in NER and RE fine-tuning tasks for biomedical datasets,demonstrating its robustness in domain specific downstream tasks. The BioMedBERT v1.0 model fine-tuned on SQuAD v2.0 dataset and further fine-tuned on BioASQ 5b, 6b and 7b datasets (trained for 2 epochs) outperforms state-of-the-art MRR scores for all 3 datasets as seen in Table 8 [1]. The best scores are highlighted in bold while the second best scores are underlined. **By this, BioMedBERT may be viewed as the new state-of-the-art results for biomedical question-answering tasks.**

Dataset	Metrics	BERT	SOTA	BioMedBERT
GAD	P	74.28	79.21	78.04
	R	85.11	89.25	82.01
	F	79.29	83.94	79.92
EU-ADR	P	75.45	81.05	76.34
	R	96.55	98.01	84.74
	F	84.62	86.51	78.43

Table 5: Results for RE fine-tuning on GAD and EU-ADR datasets.

Dataset	Metrics (Dev)	Human	BERT	BioMedBERT
SQuAD v1.1	EM	82.3	86.2	86.12
	F1	91.2	92.2	**92.46**
SQuAD v2.0	EM	86.2	78.7	**80.85**
	F1	89.5	81.9	**83.96**

Table 6: Results for QA fine-tuning on SQuAD v1.1 and v2.0 datasets.

Type	Dataset	Metrics	BERT	SOTA	BioMedBERT
Disease	NCBI Disease	P	84.12	89.04	86.02
		R	87.19	91.25	89.06
		F	85.63	89.71	87.51
	BC5CDR	P	81.97	89.61	83.79
		R	82.48	87.84	85.06
		F	82.41	87.15	84.42
Drug/chem	BC5CDR	P	90.94	94.26	92.04
		R	91.38	93.61	92.37
		F	91.16	93.47	92.21
	BC4CHEMD	P	91.19	92.8	89.43
		R	88.92	91.92	83.58
		F	90.04	92.36	86.41
Gene/Protein	BC2GM	P	81.17	85.16	81.4
		R	82.42	85.12	83.26
		F	81.79	84.72	82.32
	JNLPBA	P	69.57	74.43	70.93
		R	81.20	83.56	82.76
		F	74.94	83.56	76.39

Table 7: Results for NER fine-tuning on Biomedical NER datasets.

[1] `http://participants-area.bioasq.org/`

Dataset	Metrics	BERT	SOTA	BioMedBERT v1.0			
				SQuAD v1.1		SQuAD v2.0	
				epochs=2	epochs=5	epochs=2	epochs=5
BioASQ 4b	SAcc	27.33	**34.76**	29.99	32.92	32.59	32.30
	LAcc	44.72	**50.88**	48	50.31	48.44	50.31
	MRR	33.77	**41.34**	30.41	39.9	38.28	40.00
BioASQ 5b	SAcc	39.33	46.66	**46.8**	45.33	46.68	46.00
	LAcc	52.67	60.38	60.81	**61.33**	61.29	**61.33**
	MRR	44.27	52.12	51.65	51.85	**52.14**	51.75
BioASQ 6b	SAcc	33.54	**42.86**	41.84	39.13	41.92	**42.86**
	LAcc	51.55	61.18	61.06	60.87	**62.12**	60.25
	MRR	40.88	49.05	49.33	47.64	**50.50**	49.81
BioASQ 7b	SAcc	-	40.12	41.98	**46.3**	40.72	41.36
	LAcc	-	**61.11**	54.46	57.4	60.2	59.88
	MRR	-	48.47	46.95	**50.4**	48.64	49.03

Table 8: Results for QA fine-tuning on BioASQ datasets.

6.4 Information Retrieval Task Formulation & Experimental Results

One of the most challenging parts of our research was to create a validation framework for our end-to-end biomedical retrieval methodology. To accomplish that, we had to ensure two major things. First, the validation corpus should be biomedical in nature, as our embeddings are primarily trained on the biomedical corpus. Second, the word length distribution of the validation corpus used for retrieval should be similar to the word length distribution of the abstracts of the biomedical research papers in the BREATHE corpus in order to have a meaningful validation. Both of these constraints were satisfied by formulating the BioASQ dataset intelligently where we retrieve the 'context' from the 'question', rather than the 'answers' which are typically much smaller.

Additionally, we discovered the bias in the BioASQ dataset from a retrieval perspective, due to the high percentage of intersection between the words in the question and the context - which won't be the case in real life scenarios. Moreover, having a higher percentage intersection between questions and answers will cause results to be biased towards only Elasticsearch based approach (Figure 2).

To debias the dataset, we remove the records that have a very high percentage of common thresholds (based on the Jaccard index) (Leskovec et al., 2020). The result in Table 9 shows that our methodology significantly outperforms other models on the re-structured BioASQ dataset, which is a major novelty in our research work.

Retrieved from Elasticsearch	MRR@5				MRR@10			
	ES	ES+ BioMedBERT	ES+ Glove	ES+ BERT	ES	ES+ BioMedBERT	ES+ Glove	ES+ BERT
Top 20	0.172	**0.253**	0.100	0.125	0.20	**0.285**	0.124	0.157
Top 30	0.173	**0.237**	0.090	0.115	0.20	**0.273**	0.116	0.138
Top 40	0.173	**0.231**	0.100	0.113	0.21	**0.272**	0.111	0.132
Top 50	0.172	**0.225**	0.082	0.097	0.20	**0.269**	0.098	0.117

Table 9: MRR Evaluation on BioASQ dataset with $\leq$30% intersection.

7 Discussion

We observe that the BioMedBERT model achieves state-of-the-art results in the QA tasks for both of the datasets in the biomedical domain (see Table 8) as well as in the general language domain with SQuAD v1.1 and 2.0 (see Table 6). Additionally, the model showed robustness with an impressive

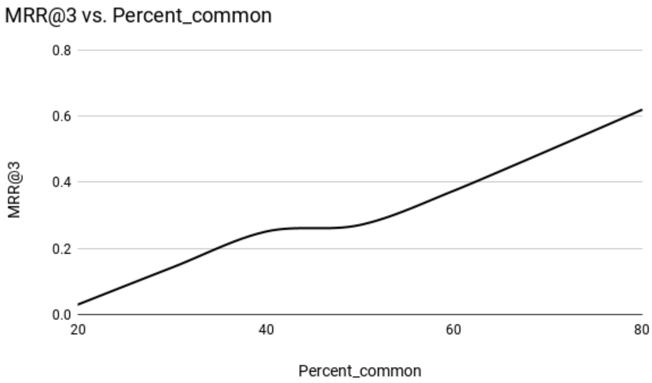

Figure 2: BioASQ Overlap and MRR Evaluation.

performance compared to BERT and the extant state-of-the-art in other language tasks such as Named Entity Recognition. While the model did not outperform the state-of-the-art, it was mostly on-par with BERT for most of the NER performance metrics. The same case for model-robustness can be made for our model's performance in Relation Extraction tasks.

Current challenges related to research into COVID-19 directed us to build a language mining tool to support question-answering and information retrieval for the biomedical domain. With respect to the main purpose of this research, **our results are the current state-of-the-art given the model performances on BioASQ question-answering datasets**. Our model outperformed numbers from BioBERT (Lee et al., 2019) on QA tasks irrespective of the fact that BioBERT was trained on a total of 18B words from the biomedical domain (13.5B from PMC full-text articles and 4.5B from PubMed abstracts), while our BioMedBERT model was trained on just over 4.1B words. The key reason for our better performance was the diversity in the datasets on which the BioMedBERT model was trained. Diversity helped in enhancing both the performance and robustness for the BioMedBERT model.

Another important and novel aspect of our work was how we framed the BioASQ dataset to validate our information retrieval methodology. Specifically, we debiased reflect reality and achieved, robust and relevant results. BioMedBERT embeddings coupled with Elasticsearch outperformed the retrieval performance of the other models based on the Mean Reciprocal Rank values as shown in Table 9. Here we experimented based on retrieving variable number of documents using Elasticsearch, and then re-ranked using embeddings to possibly eliminate any further source of bias. Our methodology (BioMedBERT + ES) outperforms the others by significant margins for all the cases.

8 Conclusion

In this paper, we present the BioMedBERT model pre-trained on the BREATHE v1.0 dataset, one of the largest and most diverse datasets of biomedical research literature. BioMedBERT achieves state-of-the-art results when fine-tuned on Question and Answering datasets, and also produces impressive performances on other language tasks such as Named Entity Recognition and Relation Extraction. BioMedBERT embeddings coupled with Elasticsearch gives state-of-the-art performance on the re-framed BioASQ dataset. Moreover, the BioMedBERT model achieves state-of-the-art results for multiple tasks even when only pre-trained on the BREATHE v1.0 dataset, which contains just over 6 million articles. Work is in progress to train an improved BioMedBERT model on the BREATHE v2 dataset with over 16 million articles. We believe continued enhancements of the BioMedBERT model will help biomedical researchers discover meaningful insights from literature faster, and make significant improvements in their field.

We would like to thank Daniel Goncharov and 42 School Silicon Valley for their contributions to building the BREATHE dataset. We would also like to thank Dave Elliott with Google Cloud Platform and Google TensorFlow Research Cloud for providing infrastructure and guidance throughout the project.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Àlex Bravo, Janet Piñero, Núria Queralt-Rosinach, Michael Rautschka, and Laura I Furlong. 2015. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC bioinformatics*, 16(1):55.

Michael J Coren. 2020. The number of scientific papers on the novel coronavirus is doubling every 14 days. *https://qz.com/*, Apr.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Maximilian Eibl and Gaedke. 2017. Evaluating the impact of word embeddings on similarity scoring in practical information retrieval. In *Gesellschaft fur Informatik*.

Daniel Goncharov, David Elliott, and Soonson Kwon. How the google ai community used cloud to help biomedical researchers — google cloud blog.

Clinton Gormley and Zachary Tong. 2015. *Elasticsearch: the Definitive Guide: A Distributed Real-time Search and Analytics Engine*. " O'Reilly Media, Inc.".

Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. 2016. Lstm: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 28(10):2222–2232.

K Hao. 2020. Over 24,000 coronavirus research papers are now available in one place.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Jin-Dong Kim, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Nigel Collier. 2004. Introduction to the bio-entity recognition task at jnlpba. In *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications*, pages 70–75. Citeseer.

Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel M Lowe, et al. 2015. The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of cheminformatics*, 7(1):1–17.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Esther Landhuis. 2016. Scientific literature: Information overload. *Nature*, 535(7612):457–458.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Jure Leskovec, Anand Rajaraman, and Jeffrey D. Ullman. 2020. *Mining of massive datasets*. Cambridge University Press.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04805*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Larry Smith, Lorraine K Tanabe, Rie Johnson nee Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christoph M Friedrich, Kuzman Ganchev, et al. 2008. Overview of biocreative ii gene mention recognition. *Genome biology*, 9(S2):S2.

Wilson L Taylor. 1953. “cloze procedure”: A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):138.

Peter Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37, 03.

Erik M Van Mulligen, Annie Fourrier-Reglat, David Gurwitz, Mariam Molokhia, Ainhoa Nieto, Gianluca Trifiro, Jan A Kors, and Laura I Furlong. 2012. The eu-adr corpus: annotated drugs, diseases, targets, and their relationships. *Journal of biomedical informatics*, 45(5):879–884.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Xuan Wang, Yu Zhang, Xiang Ren, Yuhao Zhang, Marinka Zitnik, Jingbo Shang, Curtis Langlotz, and Jiawei Han. 2019. Cross-type biomedical named entity recognition with deep multi-task learning. *Bioinformatics*, 35(10):1745–1752.

Georg Wiese, Dirk Weissenborn, and Mariana Neves. 2017. Neural domain adaptation for biomedical question answering. *arXiv preprint arXiv:1706.03610*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Wonjin Yoon, Jinhyuk Lee, Donghyeon Kim, Minbyul Jeong, and Jaewoo Kang. 2019. Pre-trained language model for biomedical question answering. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 727–740. Springer.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.