BioMegatron: Larger Biomedical Domain Language Model

Hoo-Chang Shin, Yang Zhang, Evelina Bakhturina, Raul Puri, Mostofa Patwary, Mohammad Shoeyibi, Raghav Mani
NVIDIA / Santa Clara, California, USA
hshin@nvidia.com

Abstract
There has been an influx of biomedical domain-specific language models, showing language models pre-trained on biomedical text perform better on biomedical domain benchmarks than those trained on general domain text corpora such as Wikipedia and Books. Yet, most works do not study the factors affecting each domain language application deeply. Additionally, the study of model size on domain-specific models has been mostly missing. We empirically study and evaluate several factors that can affect performance on domain language applications, such as the sub-word vocabulary set, model size, pre-training corpus, and domain transfer. We show consistent improvements on benchmarks with our larger BioMegatron model trained on a larger domain corpus, contributing to our understanding of domain language model applications. We demonstrate noticeable improvements over the previous state-of-the-art (SOTA) on standard biomedical NLP benchmarks of named entity recognition, relation extraction, and question answering.

1 Introduction
Effectively transferring the success of BERT (Devlin et al., 2018) to the biomedical domain, most notably Lee et al. (2019) (BioBERT) and Beltagy et al. (2019) (SciBERT) inspired a large number of similar works last year. For example, Peng et al. (2019); Alsentzer et al. (2019); Huang et al. (2019) added clinical text to the PubMed biomedical pre-training corpus and tested on standard biomedical and clinical NLP benchmarks. Many other similar works appeared at the ACL BioNLP Workshop (Demner-Fushman et al., 2019).

More recently, Gu et al. (2020) performed a comprehensive study on the pre-training corpus domain, language model masking method, and adversarial training, benchmarking on a number of different datasets for token classification, sequence classification, and sequence regression.

Compared to the previous works, we perform a more detailed study on (1) subword vocabulary, (2) labeling method, (2) model size, and (3) domain transfer, showing gains in token classification, sequence classification, and question answering.

2 Related Works
A prime example of Language Models (LMs) in the biomedical domain is BioBERT (Lee et al., 2019). It is a transformer LM pre-trained on the PubMed (www.ncbi.nlm.nih.gov/pubmed) biomedical text corpus comprised of biomedical literature abstracts. Their pre-training started from the checkpoint of Devlin et al. (2018) trained on Wikipedia and Books-Corpus. Independently, Beltagy et al. (2019) (SciBERT) pre-trained BERT from scratch using their vocabulary set on scientific text corpora, including PubMed abstracts and computer science papers. Both demonstrated increased performance over the previous non-BERT SOTA on biomedical benchmarks, including Named Entity Recognition (NER), Relation Extraction (RE), and Question Answering (QA). BioBERT and SciBERT report similar results on NER and RE, while only BioBERT report QA results.

They inspired other follow-up works (Alsentzer et al., 2019; Huang et al., 2019; Peng et al., 2019), most notably translating their success to the clinical domain, adding the MIMIC-III (Johnson et al., 2016) clinical text corpus. Gu et al. (2020) (PubMedBERT) used the PubMed full-text for pre-training in addition to the abstracts, and use a domain vocabulary set learned from PubMed corpus.

Meanwhile, they mostly report similar NER and RE tests and results, and only BioBERT reports QA.
results. Additionally, most use a BERT_{Base} with 110M parameters. Peng et al. (2019) report slightly improved performance on RE using BERT_{Large} while reporting worse results on NER, compared to BERT_{Base}. These results on biomedical tasks do not benefit from scaling model size to the same degree as standard NLP benchmarks such as GLUE or SQuAD (Shoeybi et al., 2019; Raffel et al., 2019).

3 Language Model Pre-training

BERT_{Base & Large} We compare our models to the pre-trained BERT_{Base & Large} models of BioBERT (Lee et al., 2019) and PubMedBERT (Gu et al., 2020) (BERT_{Base}) for fine-tuning and evaluation. For QA we use the BERT_{Large} variant of BioBERT following the authors’ recommendation.

BioMegatron Megatron-LM (Shoeybi et al., 2019) was introduced for efficient model parallel training of large LMs, with up to 8.3B parameters. Shoeybi et al. (2019) showed that rearranging the order of the layer normalization and the residual connections is critical to enabling the scaling of the BERT-style models beyond 336m parameters, and we use the same architecture.

Megatron-LM also used a larger pre-training text corpus, comprised of Wikipedia (Devlin et al., 2018), CC-Stories (Trinh and Le, 2018), RealNews (Zellers et al., 2019), and OpenWebtext (Radford et al., 2019). For our LM training, we use the 4.5 billion-word PubMed abstract set and the 1.6 billion-word CC0-licensed Commercial Use Collection of the PMC full-text corpus (www.ncbi.nlm.nih.gov/pmc).

We train three sizes of BioMegatron: with 345 million, 800 million, and 1.2 billion number of parameters (Table 1). We compare four pre-training scenarios in the smallest 345m model - using BERT-cased/uncased vocabularies, each pre-trained from scratch and fine-tuned from general domain LM. We also compare two sets of domain vocabularies learned on PubMed text corpus using SentencePiece (github.com/google/sentencepiece) library, each containing 30k and 50k subword units.

We train the larger BioMegatron models with less variation: 800m models from scratch on PubMed with BERT -cased/-uncased vocabularies; and 1.2b model starting from general domain LM checkpoint using BERT-uncased vocabulary.

| #Parameters | #Layers | #Hidden Size | #Attention Heads |
|-------------|---------|--------------|------------------|
| 345m        | 24      | 1024         | 16               |
| 800m        | 36      | 1280         | 20               |
| 1.2b        | 24      | 2048         | 16               |

Table 1: Model configurations.

4 Downstream Benchmark Tasks

We use the most widely used downstream biomedical benchmark datasets for NER, RE, and QA.

Named Entity Recognition The BC5CDR (Li et al., 2016) NER dataset annotated disease and chemical terms with IOB tagging (Ramshaw and Marcus, 1999). In NCBI-disease (Doğan et al., 2014), only disease entities are IOB-tagged.

Relation Extraction The ChemProt (Krallinger et al., 2015) dataset contains sentences from PubMed abstracts, where chemical-protein interaction types are annotated as five categories. Relation Extraction is essentially a sequence classification task, classifying a set of sentences into a category.

Question Answering The BioASQ-7b factoid task (Tsatsaronis et al., 2015) is a biomedical QA dataset whose format is similar to the SQuAD dataset (Rajpurkar et al., 2016). In this task, context-snippet, question and answer triplets, and factoid question/answers are evaluated with strict accuracy (SAcc), lenient accuracy (LAcc), and mean reciprocal rank (MRR).

5 Results and Discussion

The evaluation results on NER and RE are shown in Table 2, and QA are shown in Table 3. We perform entity-level F1 NER using the official CoNLL evaluation script translated into Python (github.com/spyysalo/conlleval.py). RE uses micro-level F1, and QA uses the BioASQ evaluation script (github.com/BioASQ/Evaluation-Measures).

5.1 Named Entity Recognition

While the NER benchmark datasets appear saturated due to the small sample size, we find that the subword vocabulary is the most critical factor. Examples of tokenization with different vocabularies are shown in Figure 1. Representing named entities as single terms is more helpful than breaking them into several subtokens. Table 4 shows the rate named entities break into sub-tokens for each benchmark training set with different sub-word vocabularies. PubMedBERT vocabulary set has a low
| Benchmark      | Model                  | #Parameters | Vocabulary                  | Prec | Rec  | F1  |
|----------------|------------------------|-------------|-----------------------------|------|------|-----|
| NER BC5CDR-chem| BioBERT 110m           | BERT-cased  | 90.0                        | 93.4 | 91.7 |
|                | PubMedBERT 110m        | PubMedBERT-vocab (30k) | 92.1                        | 93.2 | 92.6 |
|                | BioMegatron 345m       | Bio-vocab-30k | 92.1                        | 93.6 | 92.9 |
|                | BioMegatron 345m       | Bio-vocab-50k | 92.9                        | 92.0 | 92.5 |
|                | BioMegatron 800m       | BERT-cased  | 91.3                        | 92.9 | 92.1 |
|                | BioMegatron 1.2b       | BERT-uncased | 92.0                        | 90.5 | 91.3 |
| NER BC5CDR-disease| PubMedBERT 110m    | PubMedBERT-uncased (30k) | 86.2                        | 88.4 | 87.3 |
|                | BioMegatron 345m       | Bio-vocab-30k | 85.2                        | 88.8 | 87.0 |
|                | BioMegatron 345m       | Bio-vocab-50k | 86.1                        | 91.0 | 88.5 |
|                | BioMegatron 800m       | BERT-cased  | 85.8                        | 90.1 | 89.7 |
|                | BioMegatron 1.2b       | BERT-uncased | 83.8                        | 89.2 | 86.4 |
| NER NCBI-disease| PubMedBERT 110m       | PubMedBERT-uncased (30k) | 85.9                        | 87.7 | 86.8 |
|                | BioMegatron 345m       | Bio-vocab-30k | 85.6                        | 88.6 | 87.1 |
|                | BioMegatron 345m       | Bio-vocab-50k | 83.7                        | 90.4 | 87.0 |
|                | BioMegatron 800m       | BERT-cased  | 80.4                        | 68.9 | 74.3 |
|                | BioMegatron 1.2b       | BERT-uncased | 83.5                        | 90.1 | 86.7 |
| RE ChemProt     | BioBERT 110m           | BERT-cased  | 76.5                        | 73.3 | 74.8 |
|                | PubMedBERT 110m       | PubMedBERT-uncased (30k) | 73.6                        | 77.7 | 75.6 |
|                | BioMegatron 345m       | Bio-vocab-30k | 77.8                        | 72.5 | 75.1 |
|                | BioMegatron 345m       | Bio-vocab-50k | 74.5                        | 79.7 | 77.0 |
|                | BioMegatron 800m       | BERT-cased  | 80.4                        | 68.9 | 74.3 |
|                | BioMegatron 1.2b       | BERT-uncased | 83.5                        | 90.1 | 86.7 |

Table 2: Evaluation results on NER and RE after fine-tuning for 30 epochs with hyper-parameter settings of: num-fc-layers: [1, 2]; fc-hidden-size: [512, 1024]; fc-dropout: 0.5; max-seq-length: 128; learning-rate: 5e-5; cross-entropy loss, with Adam optimizer. BioMegatron models are pre-trained from scratch on PubMed, except 1.2b model which is fine-tuned from a general domain model checkpoint.

| Benchmark     | Model                  | #Parameters | Vocabulary         | SAcc | LAcc | MRR |
|---------------|------------------------|-------------|--------------------|------|------|-----|
| QA BioASQ-7b-factoid| BioBERT-Base      | 110m        | BERT-cased         | 30.8 | 64.1 | 41.1|
|               | BioBERT-Large         | 345m        | BERT-cased         | 42.8 | 62.8 | 50.1|
|               | BioMegatron           | 345m        | BERT-uncased       | 46.2 | 62.6 | 52.5|
|               | BioMegatron           | 800m        | BERT-uncased       | 45.2 | 58.6 | 50.4|
|               | BioMegatron           | 1.2b        | BERT-uncased       | 47.4 | 60.9 | 52.4|

Table 3: Evaluation results on QA after fine-tuning for 30 epochs on checkpoints fine-tuned on SQuAD dataset with fixed hyper-parameter settings as num-fc-layers: 2; fc-hidden-size: 2048; fc-dropout: 0.1; max-seq-length: 512; learning-rate: 3e-5; cross-entropy loss, using Adam optimizer. BioMegatron models are pre-trained from scratch on PubMed, except 1.2b model which is fine-tuned from a general domain model checkpoint.

Figure 1: Examples of tokenization with different sub-word vocabularies. Blue and purple text show word-level and subtoken-level entity labeling, respectively.

- **Named entity**<br>  "undifferentiated"<br>  "tachyarrhythmias"
  - BERT-cased tokenization<br>  - BioBERT-uncased (30k)<br>  - PubMedBERT-uncased (30k)

- **Sub-word vocabulary**<br>  BC5-chem: 3.012 2.42<br>  PubMedBERT-uncased (30k): 1.654 1.236<br>  BioMegatron-uncased (30k): 1.753 1.272<br>  BioMegatron-uncased (50k): 1.478 1.116

Table 4: The rate of named entities breaking into sub-tokens (#tokens/#words) in NER training sets.

We can label the entities for NER training as:<br> (1) marking the whole entity as a single label, and<br> (2) labeling sub-tokens separately. Figure 1 shows examples of the labeling methods. We find these different schemes can result in as much as ∼2% break-out rate while being smaller in size than our 50k-size vocabulary. A lower break-out rate with smaller vocabulary size probably helps achieve better NER performance despite smaller model size.
difference in the F1-score on NER evaluation, possibly indicating that the datasets are too small. We report NER results by labeling sub-tokens separately, except for NCBI-disease dataset, which results in better whole-entity labeling across models.

5.2 Relation Extraction

Since RE is a classification task, albeit on sequences rather than on tokens, the choice of sub-word vocabulary has a notable effect. We can also observe that larger models result in higher precision for lower recall, both for NER and RE. More hyper-parameter tuning could achieve higher F1-scores, even the generalization ability of such result may be questionable.

5.3 Question Answering

Table 3 show evaluation results after fine-tuning on SQuAD for 10 epochs and BioASQ for 30 epochs each, following the recipe found to work best by Lee et al. (2019). We found large batch size to be beneficial, as Q&A pairs repeat up to 88 times. We use batch size of 64 per GPU with data parallelism on 16 GPUs. Using biomedical vocabularies result in much worse results, possibly due to its low relevance in the first SQuAD fine-tuning task.

Larger models tend to perform better in QA, though it levels off after 345m parameters. The larger model size effect is more evident when fine-tuning on BioASQ directly, as shown in Table 5.

Table 5: Results on BioASQ-7b factoid, without fine-tuning on SQuAD dataset first. The other models, including those using domain vocabularies, could not achieve any comparable results. A consistent pattern of improvement over model size noticeable on par with findings in general domain LM on SQuAD.

| Model              | SAcc | LAcc | MRR |
|--------------------|------|------|-----|
| BioMegatron-345m   | 33.1 | 50.4 | 39.8|
| BioMegatron-800m   | 37.7 | 56.3 | 45.1|
| **BioMegatron-1.2b** | **40.6** | **53.7** | **45.6** |

Table 6 shows F1-score evaluation on NER and RE benchmarks using a general-domain BioMegatron-1.2b model on PubMed. Cross-domain LMs should be trained sufficiently long on domain text to achieve comparable performance.

| Benchmark         | Fine-tuning steps | F1   |
|-------------------|-------------------|------|
| SQuAD             | 10^3 steps        | 63.2 |
|                   | 10^4 steps        | 74.3 |
|                   | 10^5 steps        | 89.7 |
| BC5CDR-chem       | 2 · 10^5 steps    | 89.37|
|                   | 3 · 10^5 steps    | 91.8 |
|                   | 4 · 10^5 steps    | **92.1** |
| NER               | 5 · 10^5 steps    | 91.2 |
|                   | 10^5 steps        | 39.4 |
|                   | 10^6 steps        | 63.6 |
|                   | 10^7 steps        | 79.8 |
| BC5CDR-disease    | 2 · 10^5 steps    | 81.2 |
|                   | 3 · 10^5 steps    | 79.2 |
|                   | 4 · 10^5 steps    | **81.9** |
|                   | 5 · 10^5 steps    | 81.8 |
| RE                | 2 · 10^5 steps    | **71.1** |
|                   | 3 · 10^5 steps    | 70.4 |
|                   | 4 · 10^5 steps    | 69.7 |
|                   | 5 · 10^5 steps    | 68.3 |

Table 7 shows the results of general-domain LMs fine-tuned on BioASQ-7b-factoid. Larger models do not perform better, which may indicate overfitting is occurring on the small training set.

Table 8 shows the generalization ability of...
| Model                  | SQuAD-v1.1 | SQuAD-v2.0 |
|------------------------|-------------|-------------|
| BioMegatron-345m       | 90.4        | 84.2        |
| BioMegatron-345m-ft    | 86.5        | 77.9        |
| BioMegatron-800m       | 91.6        | 86.1        |
| BioMegatron-1.2b-ft    | 91.8        | 86.4        |
| BERTLARGE              | 90.9        | 81.8        |
| RoBERTa                | 94.6        | 89.4        |
| Megatron-3.9b          | 95.8        | 91.2        |

Table 8: Fine-tuning on SQuAD -v1.1/-v2.0 using BioMegatron and evaluating on F1-score on dev-set. BioMegatron with ‘-ft’ are pre-trained from general domain checkpoints (fine-tuned). Results of other general domain LMs are compared: RoBERTa (Liu et al., 2019), Megatron-LM (Shoeybi et al., 2019).

BioMegatron models on SQuAD datasets. Here, a large biomedical LM pre-trained on large text corpus performs better than smaller general domain LMs such as BERT\textsubscript{LARGE}, even when pre-trained on the biomedical text.

5.5 Other Domain-Specific Factors

5.5.1 Size and Bias in Biomedical Datasets

Annotating biomedical data requires in-depth domain knowledge. Besides, data often have substantial label bias as the occurrences of “abnormal” or “findings” are rare by nature. As a result, biomedical benchmark data tend to be smaller and highly biased than their general domain counterparts.

| Task | Dataset    | # Samples | Bias % |
|------|------------|-----------|--------|
| NER  | CONLL-2003 | 14987     | 0.18   |
|      | BC5CDR     | 5235      | 0.08   |
| CLS  | MRPC       | 3668      | 0.48   |
|      | ChemProt   | 19461     | 0.27   |
| QA   | SQuAD-v1.0 | 87599     | 0.4    |
|      | BioASQ-7b  | 5537      | 0.02   |

Table 9: Label bias in general and biomedical benchmark dataset. CONLL-2003 (Sang and De Meulder, 2003), MRPC (Dolan et al., 2005), and SQuAD (Rajpurkar et al., 2016) are general domain dataset for NER, CLS (RE), and QA, respectively, for comparison against biomedical domain dataset. Label bias is computed as \[\frac{\text{sum of the } \#\text{samples of minority labels}}{\text{#samples of majority label}}\] for NER and RE (CLS), and \[\frac{\text{#minimum repeat of the same answer}}{\text{#maximum repeat of the same answer}}\] for QA.

Table 9 shows a comparison of benchmark datasets for NER, RE (CLS), and QA in the biomedical domain and their general-domain counterparts. The SQuAD Q&A set is 15 times larger than the BioASQ data, where the same question-answer combinations appear up to 88 times in BioASQ. Question-answer pairs are seldom repeated in SQuAD data, at most twice. The BC5CDR NER dataset is 1/3 size of CONLL-2003 and the ratio of $I/O$ to $O$ tags 0.08, compared to 0.18 for CONLL.

Methods to circumvent data imbalance issues such as oversampling the minority classes (Chawla et al., 2002; Chen et al., 2010) and using weighted cross-entropy gave minor effects on our NER and RE benchmarks. Recently, Li et al. (2019) proposed dice-loss for data-imbalance issues in NLP, with SOTA results on NER and QA, which could be a future avenue to explore for domain LMs. Transfer learning showed effectiveness in the biomedical QA task. However, it is somewhat unclear how to apply it to NER and RE tasks.

| Model        | PubMed Corpus      | #Words |
|--------------|--------------------|--------|
| BioBERT      | abstracts          | 4.5 billion |
| PubMedBERT   | abstracts + full-text | 16.8 billion |
| BioMegatron  | abstracts + full-text-CC | 6.1 billion |

Table 10: Pre-training text corpus of each biomedical LM. We pre-train on PubMed abstracts and full-text commercial-collection (CC) that are free of copyrights.

Pre-training Corpus and Duration
PubMedBERT is pre-trained on a much larger text corpus, as shown in Table 10. It is a performant domain-LM with a larger pre-training corpus and adequate domain vocabulary compared to its model size. We pre-train our LMs for about one epoch, reaching a masked-LM loss of about 1.2 (Devlin et al., 2018). Further pre-training may be helpful, but it is challenging to have strictly controlled experiments with many different settings.

6 Conclusion

We review and test several factors that can affect the performance of domain language models. We find that a language model targeted for a domain and application performs best. For example, model size is a secondary factor to vocabulary set for token classification task. Larger model size does not necessarily translate to better performance on a cross-domain benchmark task.

This probably indicates that there is no master model that can “do it all”, at least well enough as a targeted one. The model size is a secondary factor; larger model size can probably further improve the performance of a a domain- and application-specific language model.
Acknowledgement

The authors would like to thank Sun Kim at NIH/NCBI (now at Amazon Alexa AI) for helpful discussions and suggestions.

References

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Iz Beltagy, Arman Cohan, and Kyle Lo. 2019. Scibert: Pretrained contextualized embeddings for scientific text. *arXiv preprint arXiv:1903.10676*.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Sheng Chen, Haibo He, and Edwardo A Garcia. 2010. Ramoboost: ranked minority oversampling in boosting. *IEEE Transactions on Neural Networks*, 21(10):1624–1642.

Dina Demner-Fushman, K Bretonnel Cohen, Sophia Ananiadou, and Jun’ichi Tsujii. 2019. Proceedings of the 18th bionlp workshop and shared task. In *Proceedings of the 18th BioNLP Workshop and Shared Task*.

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.

Bill Dolan, Chris Brockett, and Chris Quirk. 2005. Microsoft research paraphrase corpus. *Retrieved March*, 29(2008):63.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific natural language model pretraining for biomedical natural language processing. *arXiv preprint arXiv:2007.15779*.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.03542*.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035.
Trieu H. Trinh and Quoc V. Le. 2018. A simple method for commonsense reasoning. *CoRR*, abs/1806.02847.

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.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *CoRR*, abs/1905.12616.