On the effectiveness of small, discriminatively pre-trained language representation models for biomedical text mining

Ibrahim Burak Ozyurt
FDI Lab Dept. of Neuroscience
UCSD
La Jolla, USA
iozyurt@ucsd.edu

Abstract

Neural language representation models such as BERT (Devlin et al., 2019) have recently shown state of the art performance in downstream NLP tasks and bio-medical domain adaptation of BERT (Bio-BERT (Lee et al., 2019)) has shown same behavior on biomedical text mining tasks. However, due to their large model size and resulting increased computational need, practical application of models such as BERT is challenging making smaller models with comparable performance desirable for real word applications. Recently, a new language transformers based language representation model named ELECTRA (Clark et al., 2020) is introduced, that makes efficient usage of training data in a generative-discriminative neural model setting that shows performance gains over BERT. These gains are especially impressive for smaller models. Here, we introduce two small ELECTRA based model named Bio-ELECTRA and Bio-ELECTRA++ that are eight times smaller than BERT Base and Bio-BERT and achieves comparable or better performance on biomedical question answering, yes/no question answer classification, question answer candidate ranking and relation extraction tasks. Bio-ELECTRA is pre-trained from scratch on PubMed abstracts using a consumer grade GPU with only 8GB memory. Bio-ELECTRA++ is the further pre-trained version of Bio-ELECTRA trained on a corpus of open access full papers from PubMed Central. While, for biomedical named entity recognition, large BERT Base model outperforms Bio-ELECTRA++, Bio-ELECTRA and ELECTRA-Small++, with hyperparameter tuning Bio-ELECTRA++ achieves results comparable to BERT.

1 Introduction

Transformers based language representation learning methods such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) are becoming increasingly popular for downstream biomedical NLP tasks due to their performance advantages (Lee et al., 2019). The performance of these models comes at a steep increase in computation cost both at training and inference time. For example, we use a BERT based re-ranker as the final step in our biomedical question answering system (Ozyurt et al., 2020), where 60% of the question answering time latency is due to the BERT classifier with 110 million parameters. The increased size of the transformer models is correlated with the increased performance (Devlin et al., 2019). Since the computational cost involved at inference time for large models is a bottleneck in their practical applications in the real world especially for real time applications such as semantic search and question answering, new approaches to achieve similar performance on smaller models are getting increasingly popular. A popular approach on this end is distilling BERT to a smaller classifier such as DistillBERT (Sanh et al., 2019), TinyBERT (Jiao et al., 2019) and MobileBERT (Sun et al., 2020). However, a small and efficient model without going through the trouble of training a large model and mimicking it in a smaller model is more preferable.

BERT uses a masked language modeling (MLM) approach by masking 15% of the training sentences and learning to guess the masked tokens in a generative manner. This results BERT using only 15% of the training data. A recent approach called ELECTRA (Clark et al., 2020), introduced a new language modeling approach where a discriminative model is trained to detect whether each token in the corrupted input was replaced by a co-trained generator model sample or not. ELECTRA is computationally more efficient than BERT and outperforms BERT given the same model size, data and computation resources (Clark et al., 2020). The improvements over BERT is most impressive at

Proceedings of the First Workshop on Scholarly Document Processing, pages 104–112

https://doi.org/10.18653/v1/P17

small model sizes, which makes it an excellent candidate in pursuit of small and efficient language representation models for biomedical text mining.

In this paper, we introduce two small and efficient ELECTRA based domain-specific language representation models trained on PubMed abstracts and on PubMed Central (PMC) open-access full papers, respectively, with a domain specific vocabulary achieving comparable or (in some cases) better results on several biomedical text mining tasks to BERT Base model that have 8 times more parameters resulting in 8 times decrease in inference time. The models are trained on a modest consumer grade GPU with only 8GB RAM which is much lower bar for pre-training of domain-specific language representation models than for BERT and variants. The performance on biomedical named entity recognition (NER) of small ELECTRA models are not as impressive as in the question answering related tasks compared to BERT. However, Bio-ELECTRA++ NER performance can be significantly improved by hyperparameter tuning to achieve comparable performance to BERT.

2 Methods

2.1 Pre-training Bio-ELECTRA/Bio-ELECTRA++

Both ELECTRA and BERT are pre-trained on English Wikipedia and BooksCorpus as general purpose language models. They both also use WordPiece tokenization (Wu et al., 2016) which represents words as constructed from character n-grams of highest co-occurrence to allow out-of-vocabulary (OOV) words to be represented. Given a vocabulary size, the character n-grams (subwords) making up the vocabulary are determined from the corpus by using an objective similar to the compression algorithms to find the subwords that would generate each unique word in the corpus. OOV words are then generated by combination of subwords from the subwords vocabulary. Since the vocabulary of BERT and ELECTRA (Clark et al., 2020) are generated from general purpose corpora, a lot of biomedical domain specific words need to be composed from subwords that does not convey enough information by themselves. For example the gene BRCA1 in BERT/ELECTRA vocabulary represented as B##R##CA##1, mostly formed from single letter embedded representations. For Bio-ELECTRA, the vocabulary is generated using SentencePiece byte-pair-encoding (BPE) model (Sennrich et al., 2016) from PubMed abstract texts from 2017. Using this domain-specific vocabulary BRCA1 is represented as BRCA##1. In this case, the composition from parts conveys more information since the learned vector embedding of BRCA subword is more likely to capture, for example, its breast cancer relatedness.

19.2 million most recent PubMed abstracts (having PMID greater than 10 million) as of March 2020 are used for Bio-ELECTRA pre-training. Sentences extracted from the paper title and abstract text are used to build the pre-training corpus of about 2.5 billion words. Using the PubMed abstract corpus and 2017 PubMed abstracts generated SentencePiece vocabulary a ELECTRA-Small model (14M trainable parameters) with a maximum sequence size of 256 and batch size of 64 is pre-trained from scratch on a RTX 2070 8GB GPU in four stages for 1.8 million steps lasting 24 days. Original ELECTRA Small was trained on a V100 32GB GPU in 4 days with a batch size of 128 for one million steps. However, the distributed ELECTRA Small++(Clark et al., 2020), which was used for our comparison experiments, was trained on the XLNet (Yang et al., 2019) corpus (about 33 billion subword corpus) with maximum sequence size of 512 for 4 million steps. Since the batch size of Bio-ELECTRA is half the size of the ELECTRA Small due to our GPUs memory size, two million steps are equivalent to one million ELECTRA training steps. ELECTRA Small++ is trained four times more than Bio-ELECTRA and trained on much larger corpus.

For the second stage of pre-training, full-text papers from open access subset of PubMed Central (PMC-OAI) as of May 2020 are used. Sentences extracted from all sections except the references section of the full-length papers are used to build a 12.3 billion words corpus. Bio-ELECTRA is further pre-trained for additional 1.8 million steps using this 12.3 billion words corpus on the same RTX 2070 8GB GPU for additional 24 days. The resulting pre-trained model is called Bio-ELECTRA++ analogous to ELECTRA Small++.

2.2 Fine-tuning for Biomedical Text Mining Tasks

The syntactic and semantic language modeling information latently captured in the pre-trained weights of transformer models combined with a classification layer were found to provide state-

of-the-art results in many NLP tasks (Devlin et al., 2019; Clark et al., 2020). We fine-tune Bio-ELECTRA, Bio-ELECTRA++, ELECTRA Small++ and BERT Base for biomedical question answering, yes/no question answer classification, named entity recognition (NER), biomedical question answer candidate ranking and relation extraction tasks.

For biomedical question answering, we used BERT and ELECTRA architectures for SQuAD (Rajpurkar et al., 2016) for SQuAD v1.1. Similar to Wiese et al. and Lee et al. (Wiese et al., 2017; Lee et al., 2019), we have combined our BioASQ (Tsatsaronis et al., 2015) 8b training data generated factoid and list questions based training set with out-of-domain SQuAD v1.1 data set to increase performance over the smaller BioASQ data.

The biomedical yes/no question answer classification task is similar to sentiment (hedging for biomedical literature) detection where the polarity (positive/negative) of a candidate sentence needs to be detected in the context of a question. For ELECTRA and BERT, we have used their official codebase from GitHub slightly extended for our specific classification task.

Named entity recognition involves detection of names of biomedical entities in sentences and usually used for downstream tasks such as information extraction and question answering. For ELECTRA and Bio-ELECTRA/Bio-ELECTRA++, we have used the ELECTRA architecture for entity level tasks adapted for BIO annotation scheme. For BERT, we have used HuggingFace Transformers Python library single output layer entity classification architecture.

In biomedical question answering, after retrieving relevant documents, the sentences containing the answer need to be filtered and ranked for the end user. Given a set of answer candidate sentences per question, where the sentences answering the question are marked as relevant, the ranking problem can be cast as a 0/1 loss classification problem and the learned probability estimates can be used to rank the candidate sentences by relevance. Due to highly unbalanced nature of this data set (on average one positive example per 99 negative examples), we have also investigated a weighted loss function. This ranking approach is also compared to cosine distance based ranking on sentence embeddings generated by Sentence-BERT (Reimers and Gurevych, 2019) with and without domain adaptation. For Sentence-BERT domain adaptation, we had further trained Sentence-BERT Siamese BERT classifier model with the training portion of our ranking data.

In biomedical relation extraction, a predetermined set of relations among two biomedical entities of interest are classified. For BERT and ELECTRA, relation extraction can be cast as a sentence classification task where the biomedical entities of interest are anonymized using pre-defined tokens to indicate to the classifier the identity of the named entities are not important compared to the context.

For each fine-tuning experiment, ten randomly initialized models are trained and average testing performances and standard deviations are reported. Default BERT and ELECTRA hyperparameters including the number of epochs (two for QA task and three for classification/NER tasks) are used for corresponding experiments. More performance can be squeezed out of the fine-tuned models by hyperparameter tuning. For data sets with an explicit development set, we have investigated the effect of the hyperparameter tuning. All of the ELECTRA based fine-tuning trainings are conducted on a GTX 1060 6GB GPU, while the eight times larger BERT models required training on our RTX 2070 8GB GPU. For BERT experiments, cased BERT Base model is used.

3 Results

3.1 Datasets

For biomedical question answering and yes/no answer classification tests, we have generated training and testing data sets from the publicly available 2020 BioASQ (Tsatsaronis et al., 2015) Task B (8b) training data set. BioASQ 8b training set consists of 3243 questions together with ideal and exact answers and gold standard snippets. The questions come in four categories (i.e. factoid, list, yes/no and summary). Factoid and list questions are usually answered by a word or phrase (multiple word/phrases for list questions) making them amendable for extractive answer span detection type exact question answering for which general purpose question answering data sets are available such as SQUAD (Rajpurkar et al., 2016). Snippets matching their corresponding exact answer(s) are selected for the bio-medical question answering labeled set generation. For about 30% of the fac-

toid/list questions no snippet can be aligned with their corresponding ideal answers. We analyzed those cases and were able to recover additional 152 questions after manual inspection for synonyms and transliterations to include in our labeled data set. The labeled data set is split into 85%/15% training/testing data sets of size 9557 and 1809, respectively.

For yes/no answer classification, the ideal answer text of each BioASQ yes/no questions is used as the context and the exact answer (i.e. 'yes' or 'no') as label for binary classification. The ideal answers are cleaned up to remove the exact answer (yes or no) that sometimes occur at the beginning of the ideal answer. The labeled data is split into 85%/15% training/testing data sets of size 728 and 128, respectively. BioASQ yes/no questions are skewed towards yes answers where about 80% of the answers were 'yes'.

For named entity recognition tests, we have used publicly available datasets used by Crichton et. al (Crichton et al., 2017). Four common biomedical entity types are considered, namely disease, drug/chemical, gene/protein and species.

For our biomedical QA system, we have annotated up to 100 answer candidates per question as returned by the first answer ranker of our QA system as relevant or not (up to the first occurrence of a correct answer). The resulting annotated data set consists of a training set (44933 sentences for 492 questions) and a testing set (9064 sentences for 100 questions).

For biomedical relation extraction, we have used two datasets; GAD (Bravo et al., 2015) (a gene-disease relation dataset) and CHEMPROT (Krallinger et al., 2017) (a protein-chemical multi-relation dataset). For GAD, we have used the pre-processed version from Bio-BERT (Lee et al., 2019) Github repository. For CHEMPROT, we have adapted the pre-processed data from the Github repository of the relation extraction model described in (Lim and Kang, 2018) for our ELECTRA/Bio-ELECTRA/BERT experiments.

The datasets used in our experiments are summarized in Table 1. The datasets and source code are available on Github (`https://github.com/SciCrunch/bio_electra`). The Bio-ELECTRA models are available on Zenodo (`https://doi.org/10.5281/zenodo.3971235`).

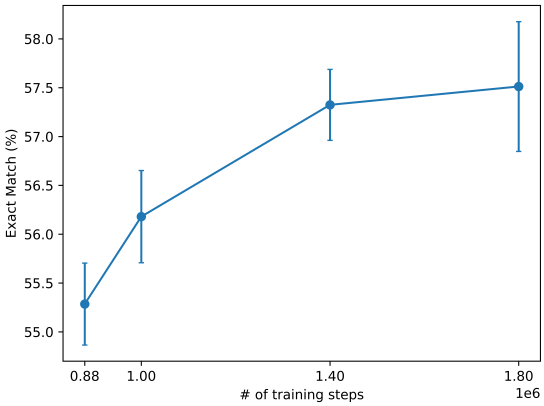

Figure 1: Change in the exact match performance for BioASQ question answering as a function of increased pre-training of Bio-ELECTRA

3.2 Effect of amount of pre-training on the Bio-ELECTRA performance

The effect of the increased number of training steps on the BioASQ question answering task is shown in Figure 1 on exact-match evaluation measure where the 95% confidence intervals are also shown, Even at 880K (or 440K in terms of ELECTRA Small++ pre-training with doubled batch size) training steps the performance of the Bio-ELECTRA is strong relative to BERT Base as shown in Table 2. Similar to what is observed in general purpose downstream question answering tasks (Devlin et al., 2019; Clark et al., 2020), more pre-training improves downstream performance in biomedical question answering.

3.3 Experimental Results

The biomedical factoid/list question answering results are shown in Table 2. We have used official SQUAD evaluation measures exact answer span match percentage and F_1 measure. While BERT Base model had slightly better performance, taken into account their 8 times smaller size and 45 times less training time (Clark et al., 2020), the performance of both Bio-ELECTRA and ELECTRA Small++ models are impressive. With one fourth of the training of ELECTRA Small++, Bio-ELECTRA has nearly same performance as the ELECTRA Small++. The best performance among ELECTRA models is observed for the Bio-ELECTRA++ model further decreasing already small performance gap between ELECTRA Small++ and BERT.

BioASQ yes/no question answer classification

Table 1: Bimedical text mining data sets

Biomedical Question Answering Dataset		
Dataset	**# training examples**	**# testing examples**
BioASQ 8b-factoid	9557	1809
Biomedical Yes/No Question Answer Classification Dataset		
Dataset	**# training examples**	**# testing examples**
BioASQ 8b-yes/no	728	128
Named Entity Recognition Datasets		
Dataset	**Entity Type**	**# training/dev/testing entities**
BC4CHEMD (Krallinger et al., 2015)	Drug/Chemical	29478/29486/25346
BC2GM (Smith et al., 2008)	Gene/Protein	15197/3061/6325
NCBI Disease (Doğan et al., 2014)	Disease	5134/787/960
LINNAEUS (Gerner et al., 2010)	Species	2119/711/1433
Biomedical Question Answer Candidate Ranking Dataset		
Dataset	**# training examples**	**# testing examples**
BioASQ 5b based	44933	9064
Relation Extraction Datasets		
Dataset	**Relation**	**# training/dev/testing examples**
GAD (Bravo et al., 2015)	Gene-disease	4796/-/534
CHEMPROT (Krallinger et al., 2017)	Protein-chemical	16521/10361/14396

Table 2: Biomedical Question Answering Test Results

Model	**Exact Match**	F_1
Bio-ELECTRA (1.8M)	57.51 (0.88)	66.87 (0.63)
Bio-ELECTRA++	57.93 (0.66)	67.48 (0.44)
ELECTRA Small++	57.78 (0.64)	67.10 (0.55)
BERT	**59.98** (0.66)	**70.25** (0.48)

task results are shown in Table 3. We have used the official BioASQ yes/no question evaluation measure of precision, recall and F_1 applied on both yes and no questions separately. While, both Bio-ELECTRA and Bio-ELECTRA++ outperforms BERT Base, BIO-ELECTRA++ is the clear winner due to its superior performance on questions with negative answer. The high standard deviations for Bio-ELECTRA and BERT Base are due to one random run in each case being stuck in a local minimum where the classifier always answers yes (since BioASQ yes/no questions are highly unbalanced towards the 'yes' answer (80% yes/20% no)).

The test results for biomedical NER experiments are shown in Table 4. Similar to BioBERT (Lee et al., 2019), we have used precision, recall and F_1 as evaluation measures. Here, the large BERT Base language representation model showed, the largest benefit over smaller models at the cost 8 times longer inference time. Bio-ELECTRA++ outperformed Bio-ELECTRA on all datasets and was better (in terms of mean F1 performance) than ELECTRA Small++ in three of the four NER entity types, while ELECTRA Small++ was slightly better than Bio-ELECTRA++ on the 'disease' entity type.

The test results for biomedical question answer candidate ranking experiments are shown in Table 5. We have used the mean reciprocal rank (MRR) to evaluate the ranking performance on the test set. Here, all of the ELECTRA models outperformed BERT, while Bio-ELECTRA++ being the best performing among them. Sentence-BERT sentence embeddings question-answer cosine similarity based approaches performed the worst.

The test results for biomedical relation extraction experiments are shown in Table 6. For the multi-relation dataset CHEMPROT, micro-averaged precision, recall and F_1 metrics are used. For GAD dataset, Bio-ELECTRA performed best closely followed by Bio-ELECTRA++. BERT showed best performance on the CHEMPROT dataset , followed by Bio-ELECTRA++.

Bio-ELECTRA++ outperformed Electra-SMALL++ in 8 out of 9 datasets spanning all five tasks. Against BERT, Bio-ELECTRA++ models showed, besides named entity recognition tasks, either competitive or better (in 3 out of 9 datasets) performance despite having only one eights of the BERT model's capacity (parameter size).

3.3.1 Effect of hyperparameter optimization on Bio-ELECTRA++

For all our BERT and ELECTRA experiments, we have used default parameters without any hyperparameter optimization. To investigate the effect of hyperparameter optimization on the test performance, we have selected the named entity datasets

Table 3: Biomedical Yes/No Question Answer Classification Test Results

Model	P (Yes)	R (Yes)	F_1 (Yes)	P (No)	R (No)	F_1 (No)
Bio-ELECTRA (1.8M)	87.99 (2.95)	**97.94** (1.35)	92.66 (1.56)	77.14 (26.47)	46.92 (16.39)	58.18 (19.91)
Bio-ELECTRA++	**91.24** (1.57)	95.29 (2.31)	**93.19** (0.75)	**78.91** (7.41)	**63.85** (7.92)	**69.84** (3.87)
ELECTRA Small++	88.18 (0.71)	94.31 (1.74)	91.14 (1.00)	69.92 (7.34)	50.38 (3.19)	58.40 (3.61)
BERT Base	87.02 (2.57)	95.49 (2.64)	90.99 (1.00)	65.15 (22.99)	43.46 (15.20)	51.71 (17.49)

Table 4: Biomedical Named Entity Recognition Test Results

Type	Dataset	Metrics	ELECTRA Small++	Bio-ELECTRA	Bio-ELECTRA++	BERT
Disease	NCBI disease	P	76.96 (0.80)	73.47 (0.92)	75.44 (1.06)	**85.43** (0.62)
		R	85.79 (0.64)	83.88 (0.64)	85.19 (0.77)	**87.08** (0.76)
		F_1	81.13 (0.69)	78.32 (0.52)	80.01 (0.68)	**86.24** (0.55)
Drug/chem.	BC4CHEMD	P	81.62 (0.53)	82.76 (0.42)	83.65 (0.18)	**91.36** (0.13)
		R	80.85 (0.47)	83.51 (0.46)	83.95 (0.27)	**89.46** (0.22)
		F_1	81.23 (0.15)	83.13 (0.18)	83.80 (0.19)	**90.40** (0.11)
Gene/protein	BC2GM	P	67.92 (0.40)	67.54 (0.48)	69.34 (0.43)	**83.95** (0.27)
		R	75.13 (0.29)	75.03 (0.16)	76.09 (0.28)	**84.30** (0.31)
		F_1	71.34 (0.27)	71.08 (0.23)	72.55 (0.30)	**84.13** (0.23)
Species	LINNAEUS	P	86.82 (1.16)	85.90 (1.53)	86.01 (1.55)	**96.01** (0.31)
		R	83.25 (1.42)	82.38 (0.72)	84.07 (0.92)	**93.90** (0.17)
		F_1	84.99 (1.02)	84.10 (0.79)	85.02 (0.59)	**94.94** (0.17)

Table 5: Biomedical Question Answer Candidate Reranking Test Results

Model	MRR
Electra Small++	0.281 (0.014)
Electra Small++ (weighted)	0.281 (0.008)
Bio-ELECTRA	0.325 (0.011)
Bio-ELECTRA (weighted)	0.332 (0.013)
Bio-ELECTRA++	**0.335** (0.017)
Bio-ELECTRA++ (weighted)	0.332 (0.013)
BERT Base	0.246 (0.007)
SBERT bert-base-nli-mean-tokens	0.181
SBERT domain-adaptation	0.163

and CHEMPROT relation extraction dataset, which have a development set to use for hyperparameter optimization. Using hyperopt (Bergstra et al., 2013) Python package, we searched for the optimum F_1 value on the corresponding development set of each dataset for the following hyperparameters; the learning rate among the values 1e-5, 5e-5, 1e-4 and 5e-4, number of epochs among the values 3, 5, 15 and 20 and batch size among the values 12, 24, 32 and 64. The best performing hyperparameter combination for each data set is then used to train ten randomly initialized Bio-ELECTRA++ based classifiers.

The test results of the effect of the hyperparameter optimization on Bio-ELECTRA++ are shown in Table 7. In all datasets, hyperparameter optimization resulted in substantial improvement over Bio-ELECTRA++ classifiers without hyperparameter optimization. For the NER datasets, the improved test performance caught up with the BERT test performance. Hyperpameter optimized bio-ELECTRA++ relation extraction classifier outperformed BERT. While BERT performance would also profit from hyperparameter optimization, BERT finetuning is more than an order of magnitude slower than Bio-ELECTRA++ finetuning impeding on its practicality.

4 Conclusion

In this paper, we have shown that small domain-specific language representation models that make more efficient use of pre-training data can achieve comparable or better (in some cases) downstream performance on several biomedical text mining tasks to BERT Base with eight times more parameters. Two domain-specific biomedical language representation models based on recently introduced ELECTRA architecture named Bio-ELECTRA and Bio-ELECTRA++ were pre-trained on a consumer grade GPU with only 8GB memory.

While, Bio-ELECTRA performance is highly competitive to BERT Base for question answering and classification tasks, its performance lags behing BERT Base for NER tasks. To further improve the performance of Bio-ELECTRA, we pre-trained it further with a second biomedical corpus of full papers from PMC open access initiative. The resulting biomedical language representation model,

Table 6: Biomedical Relation Extraction Test Results

Relation	Dataset	Metrics	ELECTRA Small++	Bio-ELECTRA	Bio-ELECTRA++	BERT
Gene-disease	GAD	P	71.06 (1.27)	72.99 (1.09)	72.48 (0.55)	**72.72** (1.08)
		R	91.35 (1.76)	92.70 (1.58)	**91.71** (1.17)	88.72 (2.12)
		F_1	79.92 (0.97)	81.66 (0.73)	**80.96** (0.35)	79.91 (1.07)
Protein-chemical	CHEMPROT	P	59.64 (2.41)	61.38 (1.90)	64.66 (1.63)	**69.75** (1.18)
		R	59.34 (2.09)	60.40 (3.12)	63.85 (2.35)	**69.87** (1.79)
		F_1	59.41 (0.88)	60.86 (2.22)	64.22 (1.40)	**69.80** (1.36)

Table 7: Effect of Hyperparameter Optimization on the Bio-ELECTRA++ Test Performance

Dataset	Metrics	Bio-ELECTRA++	Bio-ELECTRA++ opt	BERT
BC4CHEMD	P	83.65 (0.18)	88.45 (0.17)	**91.36** (0.13)
	R	83.95 (0.27)	87.44 (0.20)	**89.96** (0.22)
	F_1	83.80 (0.19)	87.94 (0.09)	**90.40** (0.11)
BC2GM	P	69.34 (0.43)	77.73 (0.38)	**83.95** (0.27)
	R	76.09 (0.28)	80.87 (0.34)	**84.30** (0.31)
	F_1	72.55 (0.30)	79.27 (0.31)	**84.13** (0.23)
NCBI disease	P	75.44 (1.06)	83.40 (0.79)	**85.43** (0.62)
	R	85.19 (0.77)	86.36 (0.65)	**87.08** (0.76)
	F_1	80.01 (0.68)	84.85 (0.65)	**86.24** (0.55)
LINNAEUS	P	86.01 (1.55)	93.77 (1.25)	**96.01** (0.31)
	R	84.07 (0.92)	**96.28** (0.65)	93.90 (0.17)
	F_1	85.02 (0.59)	**95.01** (0.84)	94.94 (0.17)
CHEMPROT	P	64.66 (1.63)	**73.23** (0.86)	69.75 (1.18)
	R	63.85 (2.35)	**71.46** (0.79)	69.87 (1.79)
	F_1	64.22 (1.40)	**72.33** (0.71)	69.80 (1.36)

Bio-ELECTRA++, outperformed Bio-ELECTRA in 8 out of 9 datasets. After hyperparameter fine-tuning, the performance lead of BERT Base over Bio-ELECTRA++ on NER tasks is drastically decreased making Bio-ELECTRA++ competitive or superior to BERT in all biomedical text mining tasks tested.

Acknowledgments

This work was supported by the NIDDK Information Network (dkNET; `http://dknet.org`) via NIHs National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) award U24DK097771.

References

J. Bergstra, D. Yamins, and D. D. Cox. 2013. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML13, page I115I123. JMLR.org.

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.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators.

G. Crichton, S. Pyysalo, and Chiu. 2017. A neural network multi-task learning approach to biomedical named entity recognition. *BMC Bioinformatics*, 18(368).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Martin Gerner, Goran Nenadic, and Casey M Bergman. 2010. Linnaeus: a species name identification system for biomedical literature. *BMC bioinformatics*, 11(1):85.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2019. Tinybert: Distilling bert for natural language understanding.

Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel M. Lowe, Roger A. Sayle, Riza Theresa Batista-Navarro, Rafal Rak, Torsten Huber, Tim Rocktäschel, Sérgio Matos, David Campos, Buzhou Tang, Hua Xu, Tsendsuren Munkhdalai, Keun Ho Ryu, SV Ramanan, Senthil Nathan, Slavko Žitnik, Marko Bajec, Lutz Weber, Matthias Irmer, Saber A. Akhondi, Jan A. Kors, Shuo Xu, Xin An, Utpal Kumar Sikdar, Asif Ekbal, Masaharu Yoshioka, Thaer M. Dieb, Miji Choi, Karin Verspoor, Madian Khabsa, C. Lee Giles, Hongfang Liu, Komandur Elayavilli Ravikumar, Andre Lamurias, Francisco M. Couto, Hong-Jie Dai, Richard Tzong-Han Tsai, Caglar Ata, Tolga Can, Anabel Usié, Rui Alves, Isabel Segura-Bedmar, Paloma Martínez, Julen Oyarzabal, and Alfonso Valencia. 2015. The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of Cheminformatics*, 7(1).

Martin Krallinger et al. 2017. Overview of the BioCreative VI chemical-protein interaction track. In *Proceedings of the BioCreative VI Workshop*, pages 141–146, Bethesda, MD.

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*, 36(4):1234–1240.

Sangrak Lim and Jaewoo Kang. 2018. Chemical-gene relation extraction using recursive neural network. *Database*, 2018. Bay060.

Ibrahim Burak Ozyurt, Anita Bandrowski, and Jeffrey S Grethe. 2020. Bio-AnswerFinder: a system to find answers to questions from biomedical texts. *Database*, 2020. Baz137.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

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, Manabu Torii, Hongfang Liu, Barry Haddow, Craig A. Struble, Richard J. Povinelli, Andreas Vlachos, William A. Baumgartner, Lawrence Hunter, Bob Carpenter, Richard Tzong-Han Tsai, Hong-Jie Dai, Feng Liu, Yifei Chen, Chengjie Sun, Sophia Katrenko, Pieter Adriaans, Christian Blaschke, Rafael Torres, Mariana Neves, Preslav Nakov, Anna Divoli, Manuel Maña-López, Jacinto Mata, and W. John Wilbur. 2008. Overview of biocreative ii gene mention recognition. *Genome Biology*, 9(2).

Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. 2020. Mobilebert: a compact task-agnostic BERT for resource-limited devices. *CoRR*, abs/2004.02984.

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 Artieres, 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*, 16:138.

Georg Wiese, Dirk Weissenborn, and Mariana Neves. 2017. Neural domain adaptation for biomedical question answering. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 281–289, Vancouver, Canada. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, ukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32*, pages 5753–5763. Curran Associates, Inc.