Pretrained Language Models for Biomedical and Clinical Tasks: Understanding and Extending the State-of-the-Art

Patrick Lewis, Myle Ott, Jingfei Du, Veslin Stoyanov

Facebook AI Research

{plewis,myleott,jingfeidu,ves}@fb.com

Abstract

A large array of pretrained models are available to the biomedical NLP (BioNLP) community. Finding the best model for a particular task can be difficult and time-consuming. For many applications in the biomedical and clinical domains, it is crucial that models can be built quickly and are highly accurate. We present a large-scale study across 18 established biomedical and clinical NLP tasks to determine which of several popular open-source biomedical and clinical NLP models work well in different settings. Furthermore, we apply recent advances in pretraining to train new biomedical language models, and carefully investigate the effect of various design choices on downstream performance. Our best models perform well in all of our benchmarks, and set new State-of-the-Art in 9 tasks. We release these models in the hope that they can help the community to speed up and increase the accuracy of BioNLP and text mining applications.

1 Introduction

The pretrain-and-finetune approach has become the dominant paradigm for NLP applications in the last few years (Peters et al., 2018; Devlin et al., 2019; Yang et al., 2019; Conneau et al., 2020, inter alia.), bringing significant performance gains in many areas of NLP. Models trained on Wikipedia and WebText (Radford et al., 2019) generally perform well on a variety of target domains, but various works have noted that pretraining on in-domain text is an effective method for boosting downstream performance further (Peters et al., 2018; Beltagy et al., 2019; Li et al., 2019; Gururangan et al., 2020). Several pretrained models are available specifically in the domain of biomedical and clinical NLP driving forward the state of the art including BioBERT (Lee et al., 2019), SciBERT (Beltagy et al., 2019), ClinicalBERT (Alsentzer et al., 2019) and BioMedRoBERTa (Gururangan et al., 2020). While it is great to have multiple options, it can be difficult to make sense of what model to use in what case – different models are often compared on different tasks. To further complicate matters, more powerful general-purpose models are being released continuously. It is unclear whether it is better to use a more powerful general-purpose model like RoBERTa, or a domain-specific model derived from an earlier model such as BioBERT. And given the opportunity to pretrain a new model, it is unclear what are the best practices to do that efficiently.

Our goal is to understand better the landscape of pretrained biomedical and clinical NLP models. To that effect, we perform a large-scale study across 18 established biomedical and clinical NLP tasks. We evaluate four popular bioNLP models using the same experimental setup. We compare them to general purpose RoBERTa checkpoints. We find that BioBERT performs best overall on biomedical tasks, but the general-purpose RoBERTA-large model performs best on clinical tasks. We then take advantage of recent advances in pretraining by adapting RoBERTa (Liu et al., 2019) to biomedical and clinical text. We investigate what choices are important in pretraining for strong downstream bioNLP performance, including model size, vocabulary/tokenization choices and training corpora. Our best models perform well across all of the tasks, establishing a new state of the art on 9 tasks. Finally, we apply knowledge distillation to train a smaller model that outperforms all other models with similar computational requirements. We will release all of our pretrained models and the scripts used to run the benchmark 18 tasks.[1]

[1] Available at https:TODO:

2 Tasks and Datasets

We select a broad range of datasets to cover both scientific and clinical textual domains, and common modelling tasks – namely i) Sequence labelling tasks, covering Named Entity Recognition (NER) and de-identification (De-id) and ii) Classification tasks, covering relation extraction, multi-class and multi-label classification and Natural Language Inference (NLI)-style tasks. These tasks were also selected to optimize overlap with previous work in the space, drawing tasks from the BLUE benchmark (Peng et al., 2019), BioBERT (Lee et al., 2019), SciBERT (Beltagy et al., 2019) and ClinicalBERT (Alsentzer et al., 2019). The tasks are summarized in Table 1 and described in the following subsections.

2.1 Sequence Labelling Tasks

BC5-CDR (Li et al., 2016) is an NER task requiring the identification of Chemical and Disease concepts from 1,500 PubMed articles. There are 5,203 and 4,182 training instances for chemicals and diseases respectively.

JNLPBA (Collier and Kim, 2004) is an NER task requiring the identification of entities of interest in micro-biology, with 2,000 training PubMed abstracts.

NCBI-Disease (Doğan et al., 2014) requires identification of disease mentions in PubMed abstracts. There are 6,892 annotations from 793 abstracts.

BC4CHEMD (Krallinger et al., 2015) requires the identification of chemical and drug mentions from PubMed abstracts. There are 84,310 annotations from 10,000 abstracts.

BC2GM (Smith et al., 2008) requires the identification of 24,583 protein and gene mentions from 20,000 sentences from PubMed.

LINNAEUS (Gerner et al., 2010) is a collection of 4,077 species annotations from 153 PubMed articles.

Species-800 (Pafilis et al., 2013) is a collection 3,708 species annotations in 800 PubMed abstracts.

I2B2-2010/VA (Uzuner et al., 2011) is made up of 871 de-identified clinical reports. The task requires labelling a variety of medical concepts in clinical text.

I2B2-2012 (Sun et al., 2013b,a) is made up of 310 de-identified clinical discharge summaries. The task requires the identification of temporal events within these summaries.

I2B2-2014 (Stubbs and Uzuner, 2015; Stubbs et al., 2015) is made up of 1,304 de-identified longitudinal medical records. The task requires the labelling of spans of text of private health information.

2.2 Classification Tasks

HOC (Baker et al., 2016) is a multi-label classification task requiring the classification of cancer concepts for PubMed Articles. We follow (Peng et al., 2019) and report abstract-level F1 score.

MedNLI (Romanov and Shivade, 2018) is a 3-class NLI dataset built from 14K pairs of sentences in the clinical domain.

ChemProt (Krallinger et al., 2017) requires classifying chemical-protein interactions from 1,820 PubMed articles. We follow the standard practice of evaluating over the 5 most common classes.

GAD (Bravo et al., 2015) is a binary relation extraction task for 5330 annotated gene-disease interactions from PubMed. We use the cross-validation splits from Lee et al. (2019).

EU-ADR (van Mulligen et al., 2012) is a small data binary relation extraction task with 355 annotated gene-disease interactions from PubMed. We use the cross-validation splits from Lee et al. (2019).

DDI-2013 (Herrero-Zazo et al., 2013) is a relation extraction task requiring recognition of drug-drug interactions. There are 4 classes to extract from 4920 sentences from PubMed, as well as many sentences which do not contain relations.

I2B2-2010-RE (Uzuner et al., 2011) in this setting of I2B2-2010, we focus on the relation extraction task to detect 8 clinical events.

3 Pretraining Corpora

There is a wide range of text corpora in the biomedical and clinical domains. We limit our options to data that is freely available to the public so that models can be open-sourced.

Task Name	Domain	Task	Metric	Task Name	Domain	Task	Metric
BC5-CDR-Chemical	PubMed	N.E.R.	F1	I2B2-2012	Clinical	N.E.R.	F1
BC5-CDR-Disease	PubMed	N.E.R.	F1	I2B2-2014	Clinical	De-ID	F1
JNLPBA	PubMed	N.E.R.	F1	HOC	PubMed	Multi-label classif.	Macro-F1
NCBI-D	PubMed	N.E.R.	F1	ChemProt	PubMed	Rel. extract.	Macro-F1
BC4CHEMD	PubMed	N.E.R.	F1	GAD	PubMed	Binary Rel. Extract.	F1
BC2GM	PubMed	N.E.R.	F1	EU-ADR	PubMed	Binary Rel. Extract.	F1
LINNEAEUS	PubMed	N.E.R.	F1	DDI-2013	PubMed	Rel. Extract.	Micro-F1
Species-800	PubMed	N.E.R.	F1	I2B2-2010-RE	Clinical	Rel. extract.	F1
I2B2-2010	Clinical	N.E.R.	F1	MedNLI	Clinical	NLI	Acc

Table 1: Summary of our considered tasks

PubMed abstracts PubMed[2] is a free resource containing over 30 million citations and abstracts of biomedical literature. PubMed abstracts are a popular choice for pretraining biomedical language models (Lee et al., 2019; Peng et al., 2020) because of the collection's large size and broad coverage. Following past work, we obtained all PubMed abstracts published as of March 2020. After removing empty abstracts we retained 27GB of text from 22 million abstracts, consisting of approximately 4.2 billion words.

PubMed Central full-text PubMed Central[3] (PMC) is an open access collection of over 5 million full-text articles from biomedical and life science research, which has been used in past scientific language modeling work (Beltagy et al., 2019). Following past work, we obtained all PubMed Central full-text articles published as of March 2020. We use the `pubmed_parser` package[4] to extract plain text from each article. After removing empty paragraphs and articles with parsing failures we retained 60GB of text from 3.4 million articles, consisting of approximately 9.6 billion words.

MIMIC-III The Medical Information Mart for Intensive Care, third update (MIMIC-III) consists of deidentified clinical data from approximately 60k intensive care unit admissions. Following related work (Zhu et al., 2018; Peng et al., 2019), we extract all physician notes resulting in 3.3GB of text and approximately 0.5 billion words.

Other corpora Other authors have used subsets of papers on Semantic Scholar (Gururangan et al., 2020; Ammar et al., 2018), but these corpora are not generally publicly available. The CORD-19 dataset (Wang et al., 2020) is a publicly-available corpus of articles focusing on COVID-19, but is largely subsumed by PMC, so we do not directly include it in our work.

4 Pretrained Models

We compare five publicly-available language models which together form a representative picture of the state-of-the-art in biomedical and clinical NLP. We use the HuggingFace Transformers library to access the model checkpoints (Wolf et al., 2019).

SciBERT (Beltagy et al., 2019) is a masked language model (MLM) pretrained from scratch on a corpus of 1.14M papers from Semantic Scholar (Ammar et al., 2018), of which 82% are in the biomedical domain. SciBERT uses a specialized vocabulary built using Sentence-Piece (Sennrich et al., 2016; Kudo, 2018)[5] on their pretraining corpus. We use the uncased SciBERT variant.

BioBERT (Lee et al., 2019) is based on the BERT-base model (Devlin et al., 2019), with additional pretraining in the biomedical domain. We use BioBERT-v1.1. This model was was trained for 200K steps on PubMed and PMC for 270K steps, followed by an additional 1M steps of training on PubMed, using the same hyperparameter settings as BERT-base.

ClinicalBERT (Alsentzer et al., 2019) is also based on BERT-base, but with a focus on clinical tasks. We use the "Bio+Clinical BERT" checkpoint, which is initialized from BioBERT, and then trained using texts from MIMIC-III for 150K steps using a batch size of 32.

RoBERTa (Liu et al., 2019) is a state-of-the-art general purpose model. We experiment with RoBERTa-base and RoBERTa-large to understand

[2] https://pubmed.ncbi.nlm.nih.gov

[3] https://www.ncbi.nlm.nih.gov/pmc

[4] https://github.com/titipata/pubmed_parser

[5] https://github.com/google/sentencepiece

how general domain models perform on biomedical tasks. Both models are pretrained with much larger batch sizes than BERT, and use dynamic masking strategies to prevent the model from overmemorization of the training corpus. RoBERTa outperforms BERT on general-domain tasks (Liu et al., 2019).

BioMed-RoBERTa (Gururangan et al., 2020) is a recent model based on RoBERTa-base. BioMed-RoBERTa is initialized from RoBERTa-base, with an additional pretraining of 12.5K steps with a batch size of 2048, using a corpus of 2.7M scientific papers from Semantic Scholar (Ammar et al., 2018).

4.1 Pretraining New Models

In addition to these publicly available models, we also pretrain new models on the corpora in Section 3 and examine which design criteria are important for strong downstream performance on BioNLP tasks. We have three criteria we are interested in studying: i) The effect of model size on downstream performance; ii) the effect of pretraining corpus on downstream performance; and, iii) whether tokenizing with a domain-specific vocabulary has a strong effect on downstream performance.

We pretrain a variety of models based on the RoBERTa-base and RoBERTa-large architectures, with detailed ablations discussed in section 6.1. We use the PubMed data, and optionally include MIMIC-III. We initialize our models with the RoBERTa checkpoints, except when we use a domain-specific vocabulary, then we retrain the model from a random initialization. Our domain-specific vocabulary is a byte-level byte-pair encoding (BPE) dictionary learned over our PubMed pretraining corpus (Radford et al., 2019; Sennrich et al., 2016). Both the general-purpose (RoBERTa) and domain-specific vocabularies contain 50k subword units. Our best performing models use PubMed abstracts, PMC and MIMIC-III pretraining and a domain-specific vocabulary, and are referred to as "ours-base" and "ours-large" in the following sections.

5 Experimental Setup

5.1 Pretraining

We largely follow the pretraining methodology of Liu et al. (2019). We pretrain models using FAIRSEQ (Ott et al., 2019) on input sequences of 512 tokens, of which 15% are masked and later predicted.[6] We pretrain with batches of 8,192 sequences and use the AdamW optimizer (Loshchilov and Hutter, 2019) with $\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 1e-6$. We regularize the model with dropout ($p = 0.1$) and weight decay ($\lambda = 0.01$). We pretrain all models for 500k steps using mixed precision on V100 GPUs. We linearly warmup the learning for the first 5% of steps and linearly decay the learning rate to 0 over the remaining steps. We use a learning rate of 6e-4 for base models and 4e-4 for large models.

5.2 Fine-tuning

We fine-tune models using 5 different seeds and report the median result on the test sets.

For sequence labelling tasks, we use learning rate of 1e-5 and a batch size of 32. For all sequence labelling tasks, we train for 20 epochs in total and choose the best checkpoint based on validation set performance (evaluating every 500 optimization steps). We fine-tuned the models with 5 seeds and report the median test results across these seeds.

For classification tasks, we use a learning rate of 0.002 and a batch size of 16. For HOC, ChemProt, MedNLI and I2B2-2010-RE, we run for a maximum of 10 epochs, and perform early stopping, evaluating performance on validation data every 200 optimization steps. As GAD and EU-ADR are split into 10 train/test cross-validation partitions, we choose early-stopping hyperparameters using one fold, and report the median test results on the other 9 folds.

6 Results

Table 2 shows our main results. The first columns show results for the general-purpose RoBERTa-base checkpoint, the next four show results for the specialized models mentioned in Section 4. The Roberta-large column shows results for the general-purpose RoBERTa-large checkpoint. The "ours-base" and "ours-large" columns refers to our proposed RoBERTa-base and RoBERTa-large sized models respectively, which were trained using PubMed and Mimic-III data and a domain-specific vocabulary. We observe the following: i) RoBERTa-large outperforms RoBERTa-base consistently, despite having access to the same training

[6] Following Devlin et al. (2019) and Liu et al. (2019), with 10% probability we randomly unmask a masked token or replace it with a random token.

Task Name	RoBERTa-base	SciBERT	BioBERT	Clinical-BERT	BioMed-RoBERTa	Ours-base	RoBERTa-large	Ours-large
BC5CDR-C.	87.3	91.9	91.9	90.6	90.3	92.9	90.8	**93.7**
BC5CDR-D.	77.6	83.6	83.3	81.3	80.6	83.8	82.3	**85.2**
JNLPBA	79.5	80.3	80.4	79.3	80.2	80.6	80.1	**81.0**
NCBI-disease	84.7	86.9	87.6	86.1	86.1	87.7	87.1	**89.0**
BC4CHEMD	88.6	91.8	92.2	90.3	89.7	92.7	90.6	**93.7**
BC2GM	82.7	85.7	85.6	83.9	84.2	87.0	85.3	**88.0**
LINNEAEUS	79.8	84.1	86.2	84.8	84.2	85.3	87.8	**88.4**
Species-800	75.8	77.8	79.2	77.4	77.3	79.6	78.3	**81.1**
I2B2-2010	83.5	86.3	86.0	86.3	85.0	88.1	87.3	**89.7**
I2B2-2012	74.9	77.6	77.6	78.0	76.4	79.5	78.3	**80.8**
I2B2-2014	95.6	95.2	94.7	94.6	95.2	95.5	95.8	**96.3**
HOC	86.0	84.7	86.6	86.2	**86.7**	86.5	85.2	86.6
ChemProt	69.6	69.7	73.9	68.5	75.7	75.4	71.7	**76.2**
GAD	79.4	78.7	81.2	79.2	81.6	**82.2**	73.4	81.1
EU-ADR	85.0	**85.5**	85.0	85.1	85.0	85.0	85.0	85.0
DDI-2013	79.0	79.1	79.9	77.3	80.7	81.0	80.5	**82.1**
I2B2-2010-RE	72.4	69.8	74.4	74.0	75.0	75.0	75.2	**78.6**
MedNLI	81.4	79.7	82.5	81.8	85.1	87.1	83.3	**88.5**
Mean (Seq. Lab.)	82.7	85.6	85.9	84.8	84.5	86.6	85.8	**87.9**
Mean (Classif.)	79.0	78.2	80.5	78.9	81.4	81.7	79.2	**82.6**
Mean (PubMed)	81.1	83.1	84.1	82.3	83.3	84.6	82.9	**85.5**
Mean (Clinical)	81.6	81.7	83.0	82.9	83.3	85.1	84.0	**86.8**
Mean (all)	81.3	82.7	83.8	82.5	83.3	84.7	83.2	**85.8**

Table 2: Test results on all tasks for our RoBERTa baselines, publicly available models and our best Large and Base-sized models. All results are the median of 5 runs with different seeds

corpora; ii) We find that BioBERT performs best from the publicly available models that we experiment with; and iii) our newly introduced models perform well, achieving the best results for 17 out of the 18 tasks in our experiments, often by a large margin. The exception is EU-ADR, which has a small test set where all models achieve essentially the same classification accuracy.

Digging deeper, we note that standard RoBERTa-large is competitive with the four specialized models on sequence labelling tasks (85.8 vs 85.9) and outperforms them on clinical tasks (84.0 vs 83.3), despite having no specialized biomedical or clinical pretraining. This suggests that larger, more powerful general-purpose models could be a good default choice compared to smaller, less powerful domain-specific models. Nevertheless, applying domain-specific training to otherwise-comparable models results in significant performance gains in our experiments, as shown by comparing ours-base and ours-large to RoBERTa-base and RoBERTa-large in Table 2, (+3.5% and +2.6% mean improvement), consistent with findings from previous work (Gururangan et al., 2020).

6.1 Ablations

The "ours-base" and "ours-large" models shown in Table 2 refer to the best language models that we trained in our experiments described in Section 4.1. These models use the RoBERTa architectures, are initialized with random weights, use a BPE vocabulary learnt from PubMed, and are pretrained on both our PubMed and MIMIC-III corpora. We performed a detailed ablation study to arrive at these models, and in what follows, we analyse the design decisions in detail. A summary of these results are shown in Table 3, a description of task groupings in Table 4, and full results can be found in Appendix A.1.

6.1.1 Effect of vocabulary

The effect of learning a dedicated biomedical vocabulary for base and large models can be analysed by comparing row 2 to row 3, row 4 to 5, and row 7 to 8 in Table 3. A dedicated vocabulary consistently improves sequence labelling tasks, improving results for base models by 0.7% and our large model by 0.6% on average. The difference is less consistent for classification tasks, improving the large model by 0.5%, but reducing performance on the small model by 0.7%. A specialized domain-specific vocabulary was also shown to be

Model \ Mean	Clinical	PubMed	Seq. Lab.	Classif.	All
(1) RoBERTa-base	81.6	81.1	82.7	79.0	81.3
(2) +PM	83.5	84.1	85.7	81.1	83.9
(3) +PM+Voc.	83.4	84.4	86.5	80.4	84.1
(4) +PM+M3	85.0	84.0	85.9	81.6	84.2
(5) +PM+M3+Voc.	85.1	84.6	86.6	81.8	84.7
(6) RoBERTa-large	84.0	82.9	85.8	79.2	83.2
(7) +PM+M3	85.7	85.1	87.3	82.1	85.3
(8) +PM+M3+Voc.	**86.8**	**85.5**	**87.9**	**82.6**	**85.8**

Table 3: Ablation test set results. Rows 5 and 8 correspond to "ours-base"' and "ours-large" in Table 2 respectively. **Bold** indicates the best model overall, Underlined indicates the best base model. "PM" indicates training with PubMed and PMC corpora and "M3" refers to the MIMIC-III corpus. "Voc" indicates using a dedicated biomedical vocabulary. Details of the tasks incuded in each column are given in Table 4

Task group	Tasks in group
Clinical	I2B2-2010, I2B2-2012, I2B2-2014, I2B2-2010-RE, MedNLI
PubMed	BC5CDR-C, BC5CDR-D, JNLPBA, NCBI-D, BC4CHEMD, BC2GM, Linneaus, Species-800, HOC, ChemProt, GAD, EU-ADR, DDI-2013
Seq. Lab.	BC5CDR-C, BC5CDR-D, JNLPBA, NCBI-D, BC4CHEMD, BC2GM, Linneaus, Species-800, I2B2-2010, I2B2-2012, I2B2-2014
Classif.	HOC, ChemProt, GAD, EU-ADR, DDI-2013, I2B2-2010-RE, MedNLI

Table 4: High level grouping of tasks. "Clinical" indicates clinical tasks, "PubMed" indicates tasks based on PubMed, "Seq. Lab." refers to sequence labelling tasks, i.e. N.E.R and De-ID. "Classif." refers to classification tasks, i.e relation extraction, multi-label classification and NLI.

useful in Beltagy et al. (2019). Since our specialized vocabulary models are trained from scratch only on biomedical data, we see that Wikipedia and WebText (Radford et al., 2019) pretraining is not necessary for strong performance.

6.1.2 Effect of training corpora

Table 3 also shows the results of text corpora. Rows 1 and 2 show that, unsurprisingly, including PubMed pretraining improves results over a RoBERTa-only model, by 2.6%. Comparing row 2 to row 4 and row 3 to 5 shows that including MIMIC-III in pretraining results in a large improvement on clinical tasks over PubMed-only models (+1.5% and +1.7%) but has little effect on PubMed-based tasks (-0.1% and +0.1%).

6.1.3 Effect of model size

Consistent with findings from the recent literature (Devlin et al., 2019; Liu et al., 2019; Radford et al., 2019; Brown et al., 2020), we find that large models perform consistently better than comparable smaller ones. Comparing row 1 to row 6, row 4 to 7, and row 5 to 8 in Table 3 shows average improvements of 2%, 1.6% and 0.9% respectively. These improvements are mostly driven by improved sequence labelling performance for large models.

6.2 Comparisons to the state-of-the-art

The focus of this paper was not a state-of-the-art on specific tasks, so we prioritized consistent hyperparameter search and did not consider task-specific tuning. Nevertheless, the models that we trained compare favorably to the state-of-the-art. Table 5 shows the best results obtained for each task in our experiments. In some cases, models used in our experiments have been reported with higher results in the literature. We attribute such difference to variance in test performance, small differences in pre-processing and differing levels of hyperparameter optimization and tuning. We control for test-set variance by running each model 5 times with different random seeds and reporting median results. We also use standard hyperparameter settings as reported in the literature. Table 5 compares our results to numbers reported in the literature. The best model in our experiments sets a new State-of-the-Art in 9 out of 18 tasks, and comes within 0.1% of the best reported result in another 3 tasks.

7 Distillation

In Section 6.1.3, we noted that larger models result in better accuracy. However, they also require more computational resources to run, limiting their applicability. Recent work addresses this issue by distilling larger models into smaller ones while retaining strong performance. In this section we investigate whether distillation works well in the Bio-NLP space.

7.1 Distillation Technique

Knowledge distillation (Hinton et al., 2015) aims to transfer the performance from a more accurate and computationally expensive *teacher model* into a more efficient *student model*. Typically, the student network is trained to mimic the output distribution

Task Name	State-of-the-Art Method	Score	Our best	Task Name	State-of-the-Art Method	Score	Our best
BC5CDR-C.	Lee et al. (2019)	93.5	**93.7**	I2B2-2012	Si et al. (2019)	**80.9**	80.8
BC5CDR-D.	Lee et al. (2019)	**87.2**	85.2	I2B2-2014	Lee et al. (2019)	93.0	**96.3**
JNLPBA	Yoon et al. (2019)	78.6	**81.0**	HOC	Peng et al. (2019)	**87.3**	87.2*
NCBI-disease	Lee et al. (2019)	**89.4**	89.0	ChemProt	Lee et al. (2019)	**76.5**	76.4*
BC4CHEMD	Lee et al. (2019)	92.4	**93.7**	GAD	Bhasuran et al. (2018)	**83.9**	82.2‡
BC2GM	Lee et al. (2019)	84.7	**88.0**	EU-ADR	Lee et al. (2019)	**86.5**	85.5†
LINNEAEUS	Giorgi and Bader (2018)	**93.5**	88.4	DDI-2013	Peng et al. (2020)	81.0	**82.1**‡
Species-800	Lee et al. (2019)	75.3	**81.1**	I2B2-2010-RE	Peng et al. (2019)	76.4	**78.6**
I2B2-2010	Si et al. (2019)	**90.3**	89.7	MedNLI	Peng et al. (2020)	84.2	**88.5**

Table 5: Our best models compared to best reported results in the literature. The best model in our experiments unless otherwise stated is RoBERTa-large with PubMed, MIMIC-III and specialized vocabulary ("ours-large" in Table 2). Other models are indicated by: (*) RoBERTa-large + PubMed + MIMIC-III; (†) SciBERT; (‡) RoBERTa-base + PubMed + MIMIC-III + vocab.

or internal activations of the teacher network, while keeping the teacher network's weights fixed.

In NLP, prior work has exploring distilling larger BERT-like models into smaller ones. Most of this work trains the student network to mimic a teacher that has already been finetuned for a specific task, i.e., *task-specific distillation* (Tsai et al., 2019; Turc et al., 2019; Sun et al., 2020). Recently, Sanh et al. (2020) showed that it is also possible to distill BERT-like models in a task-agnostic way by training the student to mimic the teacher's outputs and activations on the pretraining objective, i.e., masked language modeling (MLM). Task-agnostic distillation is appealing because it enables the distilled student model to be applied to a variety of downstream tasks. Accordingly, we primarily explore *task-agnostic distillation* in this work.

Recent work has also shown the importance of student network initialization. For example, Sanh et al. (2020) find that initializing the student network with a subset of layers from the teacher network outperforms random initialization; unfortunately this approach constrains the student network to the same embedding and hidden dimension as the teacher. Turc et al. (2019) instead advocate initializing the student model via standard MLM pretraining, finding that it outperforms the layer subset approach. Unfortunately, they only consider task-specific distillation, where the teacher network has already been finetuned to the end task, reducing the generality of the resulting student network.

We combine the approaches from Sanh et al. (2020) and Turc et al. (2019) by initializing the student network via standard MLM pretraining and then performing task-agnostic distillation by training the student to mimic a pretrained teacher on the MLM objective. We use our pretrained base model as the student network and large model as the teacher network. We also experiment with aligning the hidden states of the teacher's and student's last layer via a cosine embedding loss (Sanh et al., 2020). Since our student and teacher networks have different hidden state sizes, we learn a linear projection from the student's hidden states to the dimension of the teacher's hidden states prior to computing this loss.

We distill each student for 50k steps. Similar to pretraining (Section 5.1), we distill with a batch size of 8,192 and linearly warmup the learning rate for the first 5% of steps. We use a learning rate of 5e-4 and largely follow the distillation hyperparameter choices of Sanh et al. (2020). In particular, our loss function is a weighted combination of the original MLM cross entropy loss (with a weight $\alpha_{MLM} = 5.0$), a KL divergence loss term encouraging the student to match the teacher's outputs (with a weight $\alpha_{KL} = 2.0$) and optionally a cosine embedding loss term to align the student's and teacher's last layer hidden states (with a weight $\alpha_{cos} = 1.0$). For the KL loss we additionally employ a temperature of 2.0 to smooth the teacher's output distribution, following Sanh et al. (2020) and originally advocated by Hinton et al. (2015).

7.2 Distillation Results

Results for distillation are shown in Table 6. Since distillation trains the student for an additional 50k steps, we also include a baseline that just trains the student (base) model for longer without any distillation loss terms ("ours-base + train longer").

We find that distillation only slightly outperforms the original base model (+0.2% on average) and the original base model trained longer (+0.1% on average). Aligning the student and teacher hid-

Model \ Mean	Clin -ical	Pub Med	Seq. Lab.	Classif.	All
ours-base	85.1	84.6	86.6	81.8	84.7
+ train longer	85.0	84.7	86.6	81.8	84.8
ours-large	86.8	85.5	87.9	82.6	85.8
Distillation results (teacher = large; student = base)					
distill	85.1	84.8	86.8	81.9	84.9
distill + align	85.2	84.9	86.9	81.9	85.0

Table 6: Distillation results in context with our base and large models. Distillation outperforms both the original base model and the base model trained longer. Aligning the student and teacher's hidden states further improves performance, but the best student underperforms the large (teacher) model.

den states via a cosine embedding loss brings additional albeit slight gains (+0.1% on average relative to the "distill" model). This result is consistent with findings from Turc et al. (2019) showing that pretrained student models are a competitive baseline. The best student ("distill + align") improves upon the base model (+0.3% on average) but underperforms the large teacher (-0.8% on average).

8 Related Work

Pretrained word representations have been used in NLP modelling for many years (Mikolov et al., 2013; Pennington et al., 2014; Bojanowski et al., 2016), and have been specialised for BioNLP applications (Chiu et al., 2016; Wang et al., 2018b; Zhang et al., 2019). More recently, contextual embeddings have led to robust improvements across most NLP tasks, notably, ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019), followed more recently by models such as XLNet (Yang et al., 2019), RoBERTa (Liu et al., 2019), XLM and XLM-RoBERTa (Lample and Conneau, 2019; Conneau et al., 2020) amongst others.

Several efforts adapt such models to scientific and biomedical domains. Four such models – SciBERT (Beltagy et al., 2019), BioBERT (Lee et al., 2019), ClinicalBERT (Alsentzer et al., 2019) and BioMed-RoBERTA (Gururangan et al., 2020) – are extensively covered in Section 4. Other models include BlueBERT (Peng et al., 2019), which continue to pretrain BERT checkpoints with data from PubMed and MIMIC-III. Zhu et al. (2018) and Si et al. (2019) train ELMo and BERT models on clinical data respectively. In concurrent work, Gu et al. (2020) train contextual language models for PubMed-like text, but do not consider clinical text.

Methods for training or finetuning models on downstream tasks is also an active area of research. In this paper, we focus on well-established single-task finetuning techniques for BERT-like models using standard hyperparameter settings. Si et al. (2019) use complex task-specific models to yield strong results on clinical tasks, and Peng et al. (2020) investigate STILTS techniques (Phang et al., 2019) on a suite of biomedical tasks, achieving some gains over their baselines.

In this work, we build a suite of 18 tasks to evaluate our models. Aggregated benchmarks have become a common tool in NLP research, popularized by the GLUE benchmark (Wang et al., 2018a) for language understanding and its successor SuperGLUE (Wang et al., 2019). Evaluating on a suite of tasks is common in BioNLP too. Lee et al. (2019) evaluate on a set of 15 tasks, Peng et al. (2019) evaluate on 10 tasks referred to as "BLUE", Beltagy et al. (2019) and Gururangan et al. (2020) evaluate on 7 and 2 biomedical tasks respectively. Unfortunately, often there is little overlap between efforts, and different metrics and dataset splits are often used, making cross-model comparisons challenging, hence our efforts to evaluate all these models on a single testbed. In concurrent work, Gu et al. (2020) also note this problem, and release a similar suite of tasks to ours, referred to as BLURB, but do not include clinical tasks. We plan to evaluate our language models on the "BLURB" benchmarks in future work .

9 Conclusion

In this work, we have thoroughly evaluated 6 open-source language models on 18 biomedical and clinical tasks. Of these models, we found that BioBERT was the best on biomedical tasks, but general-purpose RoBERTa-large performed best on clinical tasks. We then pretrained 6 of our own large-scale specialized biomedical and clinical language models. We determined that the most effective models were larger, used a dedicated biomedical vocabulary and included both biomedical and clinical pretraining. These models outperform all the other models in our experiments by a wide margin. Finally, we demonstrate that our base model can be further improved by knowledge distillation from our large model, although there remains a gap between the distillation-improved base model and our large model.

Acknowledgments

The authors would like to thank Jinhyuk Lee, Kyle Lo, Yannis Papanikolaou, Andrea Pierleoni, Daniel O'Donovan and Sampo Pyysalo for their feedback and comments.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly Available Clinical BERT Embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew Peters, Joanna Power, Sam Skjonsberg, Lucy Wang, Chris Wilhelm, Zheng Yuan, Madeleine van Zuylen, and Oren Etzioni. 2018. Construction of the Literature Graph in Semantic Scholar. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 84–91, New Orleans - Louisiana. Association for Computational Linguistics.

Simon Baker, Ilona Silins, Yufan Guo, Imran Ali, Johan Högberg, Ulla Stenius, and Anna Korhonen. 2016. Automatic semantic classification of scientific literature according to the hallmarks of cancer. *Bioinformatics*, 32(3):432–440. Publisher: Oxford Academic.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. 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 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Balu Bhasuran, Jeyakumar Natarajan, and . . 2018. Automatic extraction of gene-disease associations from literature using joint ensemble learning. *PloS One*, 13(7):e0200699.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information. *arXiv:1607.04606 [cs]*. ArXiv: 1607.04606.

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*, 16:55.

Tom B. Brown, Benjamin Pickman Mann, Nick Ryder, Melanie Subbiah, Jean Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, G. Krüger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric J Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Billy Chiu, Gamal Crichton, Anna Korhonen, and Sampo Pyysalo. 2016. How to Train good Word Embeddings for Biomedical NLP. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 166–174, Berlin, Germany. Association for Computational Linguistics.

Nigel Collier and Jin-Dong Kim. 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 (NLPBA/BioNLP)*, pages 73–78, Geneva, Switzerland. COLING.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

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*, 11:85.

John M. Giorgi and Gary D. Bader. 2018. Transfer learning for biomedical named entity recognition with neural networks. *Bioinformatics (Oxford, England)*, 34(23):4087–4094.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *arXiv:2007.15779 [cs]*. ArXiv: 2007.15779.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. *arXiv:2004.10964 [cs]*. ArXiv: 2004.10964.

María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. 2013. The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of Biomedical Informatics*, 46(5):914–920.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

Martin Krallinger, Obdulia Rabal, Saber Ahmad Akhondi, Martín Pérez Pérez, Jésús López Santamaría, Gael Pérez Rodríguez, Georgios Tsatsaronis, Ander Intxaurrondo, José Antonio Baso López, Umesh Nandal, Erin M. van Buel, A. Poorna Chandrasekhar, Marleen Rodenburg, Astrid Lægreid, Marius A. Doornenbal, Julen Oyarzábal, Anália Lourenço, and Alfonso Valencia. 2017. Overview of the BioCreative VI chemical-protein interaction Track.

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):S2.

Taku Kudo. 2018. Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. *arXiv:1804.10959 [cs]*. ArXiv: 1804.10959.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual Language Model Pretraining. *arXiv:1901.07291 [cs]*. ArXiv: 1901.07291.

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. _eprint: https://academic.oup.com/bioinformatics/article-pdf/36/4/1234/32527770/btz682.pdf.

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: The Journal of Biological Databases and Curation*, 2016.

Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y.-Lan Boureau, Kyunghyun Cho, and Jason Weston. 2019. Don't Say That! Making Inconsistent Dialogue Unlikely with Unlikelihood Training. *arXiv:1911.03860 [cs]*. ArXiv: 1911.03860.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA. Curran Associates Inc. Event-place: Lake Tahoe, Nevada.

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.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Evangelos Pafilis, Sune P. Frankild, Lucia Fanini, Sarah Faulwetter, Christina Pavloudi, Aikaterini Vasileiadou, Christos Arvanitidis, and Lars Juhl Jensen. 2013. The SPECIES and ORGANISMS Resources for Fast and Accurate Identification of Taxonomic Names in Text. *PloS One*, 8(6):e65390.

Yifan Peng, Qingyu Chen, and Zhiyong Lu. 2020. An Empirical Study of Multi-Task Learning on BERT for Biomedical Text Mining. *arXiv:2005.02799 [cs]*. ArXiv: 2005.02799.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher 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, Doha, Qatar. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv:1802.05365 [cs]*. ArXiv: 1802.05365.

Jason Phang, Thibault Févry, and Samuel R. Bowman. 2019. Sentence Encoders on STILTs: Supplementary Training on Intermediate Labeled-data Tasks. *arXiv:1811.01088 [cs]*. ArXiv: 1811.01088.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from Natural Language Inference in the Clinical Domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596, Brussels, Belgium. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv:1910.01108 [cs]*. ArXiv: 1910.01108.

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.

Yuqi Si, Jingqi Wang, Hua Xu, and Kirk Roberts. 2019. Enhancing Clinical Concept Extraction with Contextual Embeddings. *Journal of the American Medical Informatics Association*, 26(11):1297–1304. ArXiv: 1902.08691.

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 Suppl 2:S2.

Amber Stubbs, Christopher Kotfila, and Ozlem Uzuner. 2015. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1. *Journal of biomedical informatics*, 58(Suppl):S11–S19.

Amber Stubbs and Ozlem Uzuner. 2015. Annotating longitudinal clinical narratives for de-identification: the 2014 i2b2/UTHealth Corpus. *Journal of biomedical informatics*, 58(Suppl):S20–S29.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013a. Annotating temporal information in clinical narratives. *Journal of Biomedical Informatics*, 46 Suppl:S5–12.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013b. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *Journal of the American Medical Informatics Association : JAMIA*, 20(5):806–813.

Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. 2020. MobileBERT: a compact task-agnostic BERT for resource-limited devices. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2158–2170, Online. Association for Computational Linguistics.

Henry Tsai, Jason Riesa, Melvin Johnson, Naveen Arivazhagan, Xin Li, and Amelia Archer. 2019. Small and practical BERT models for sequence labeling. 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 3632–3636, Hong Kong, China. Association for Computational Linguistics.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-Read Students Learn Better: On the Importance of Pre-training Compact Models. *arXiv:1908.08962 [cs]*. ArXiv: 1908.08962.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association : JAMIA*, 18(5):552–556.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d\textquotesingle Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 3261–3275. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018a. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex Wade, Kuansan Wang, Nancy Xin Ru Wang, Chris Wilhelm, Boya Xie, Douglas Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The COVID-19 Open Research Dataset. *arXiv:2004.10706 [cs]*. ArXiv: 2004.10706.

Yanshan Wang, Sijia Liu, Naveed Afzal, Majid Rastegar-Mojarad, Liwei Wang, Feichen Shen, Paul Kingsbury, and Hongfang Liu. 2018b. A comparison of word embeddings for the biomedical natural language processing. *Journal of Biomedical Informatics*, 87:12–20.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

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*, pages 5754–5764.

Wonjin Yoon, Chan Ho So, Jinhyuk Lee, and Jaewoo Kang. 2019. CollaboNet: collaboration of deep neural networks for biomedical named entity recognition. *BMC Bioinformatics*, 20(10):249.

Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. 2019. BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Scientific Data*, 6(1):52. Number: 1 Publisher: Nature Publishing Group.

Henghui Zhu, Ioannis Ch Paschalidis, and Amir Tahmasebi. 2018. Clinical Concept Extraction with Contextual Word Embedding. *arXiv:1810.10566 [cs]*. ArXiv: 1810.10566.

A Appendices

A.1 Full Results

Table 7 shows all test results from our experiments.

Task Name	General-purpose		Specialized				Ours-base				Ours-Large		Ours-Distilled
	RoBERTa-base	RoBERTa-large	SciBERT	BioBERT	Clinical-BERT	BioMed-RoBERTa	+ PM	+ PM + Voc.	+ PM + M3	+ PM + M3 + Voc.	+ PM + M3	+ PM + M3 + Voc.	+ PM + M3 + Voc.
BC5CDR-C.	87.3	90.8	90.3	91.9	91.9	90.6	91.7	93.0	92.0	92.9	93.3	93.7	93.1
BC5CDR-D.	77.6	82.3	80.6	83.6	83.3	81.3	81.9	83.9	82.4	83.8	84.8	85.2	84.0
JNLPBA	79.5	80.1	80.2	80.3	80.4	79.3	80.4	80.9	80.5	80.6	80.8	81.0	80.9
NCBI-disease	84.7	87.1	86.1	86.9	87.6	86.1	88.0	88.2	87.2	87.7	88.6	89.0	88.2
BC4CHEMD	88.6	90.6	89.7	91.8	92.2	90.3	92.6	92.6	92.5	92.7	93.2	93.7	92.8
BC2GM	82.7	85.3	84.2	85.7	85.6	83.9	86.6	87.2	86.3	87.0	87.3	88.0	87.5
LINNEAEUS	79.8	87.8	84.2	84.1	86.2	84.8	85.0	85.9	82.9	85.3	87.3	88.4	85.8
Species-800	75.8	78.3	77.3	77.8	79.2	77.4	78.4	79.5	78.2	79.6	79.5	81.1	79.9
I2B2-2010	83.5	87.3	85.0	86.3	86.0	86.3	86.0	86.7	88.2	88.1	89.2	89.7	88.5
I2B2-2012	74.9	78.3	76.4	77.6	77.6	78.0	77.3	78.4	79.4	79.5	80.2	80.8	80.0
I2B2-2014	95.6	95.8	95.2	95.2	94.7	94.6	94.9	95.0	95.5	95.5	96.2	96.3	95.3
HOC	86.0	85.2	86.7	84.7	86.6	86.2	86.5	85.9	85.8	86.5	87.2	86.6	86.9
ChemProt	69.6	71.7	75.7	69.7	73.9	68.5	74.9	72.4	74.7	75.4	76.4	76.2	75.7
GAD	79.4	73.4	81.6	78.7	81.2	79.2	81.8	81.1	81.9	82.2	81.4	81.1	82.3
EU-ADR	85.0	85.0	85.0	85.5	85.0	85.1	85.0	85.0	85.0	85.0	85.0	85.0	85.0
DDI-2013	79.0	80.5	80.7	79.1	79.9	77.3	80.5	81.9	82.1	81.0	81.8	82.1	81.2
I2B2-2010-RE	72.4	75.2	75.0	69.8	74.4	74.0	74.5	73.8	76.2	75.0	75.6	78.6	75.5
MedNLI	81.4	83.3	85.1	79.7	82.5	81.8	84.7	83.2	85.7	87.1	87.2	88.5	86.8
Mean (Seq. Lab.)	82.7	85.8	84.5	85.6	85.9	84.8	85.7	86.5	85.9	86.6	87.3	87.9	86.9
Mean (Classif.)	79.0	79.2	81.4	78.2	80.5	78.9	81.1	80.4	81.6	81.7	82.1	82.6	81.9
Mean (PubMed)	81.1	82.9	83.3	83.1	84.1	82.3	84.1	84.4	84.0	84.6	85.1	85.5	84.9
Mean (Clinical)	81.6	84.0	83.3	81.7	83.0	82.9	83.5	83.4	85.0	85.1	85.7	86.8	85.2
Mean (all)	81.3	83.2	83.3	82.7	83.8	82.5	83.9	84.1	84.2	84.7	85.3	85.8	85.0

Table 7: Test set results for all our experiments. Results are medians across 5 seeds.