Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets

Yifan Peng   Shankai Yan   Zhiyong Lu
National Center for Biotechnology Information
National Library of Medicine, National Institutes of Health
Bethesda, MD, USA
`{yifan.peng, shankai.yan, zhiyong.lu}@nih.gov`

Abstract

Inspired by the success of the General Language Understanding Evaluation benchmark, we introduce the Biomedical Language Understanding Evaluation (BLUE) benchmark to facilitate research in the development of pre-training language representations in the biomedicine domain. The benchmark consists of five tasks with ten datasets that cover both biomedical and clinical texts with different dataset sizes and difficulties. We also evaluate several baselines based on BERT and ELMo and find that the BERT model pre-trained on PubMed abstracts and MIMIC-III clinical notes achieves the best results. We make the datasets, pre-trained models, and codes publicly available at `https://github.com/ncbi-nlp/BLUE_Benchmark`.

1   Introduction

With the growing amount of biomedical information available in textual form, there have been significant advances in the development of pre-training language representations that can be applied to a range of different tasks in the biomedical domain, such as pre-trained word embeddings, sentence embeddings, and contextual representations (Chiu et al., 2016; Chen et al., 2019; Peters et al., 2017; Lee et al., 2019; Smalheiser et al., 2019).

In the general domain, we have recently observed that the General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2018a) has been successfully promoting the development of language representations of general purpose (Peters et al., 2017; Radford et al., 2018; Devlin et al., 2019). To the best of our knowledge, however, there is no publicly available benchmarking in the biomedicine domain.

To facilitate research on language representations in the biomedicine domain, we present the Biomedical Language Understanding Evaluation (BLUE) benchmark, which consists of five different biomedicine text-mining tasks with ten corpora. Here, we rely on preexisting datasets because they have been widely used by the BioNLP community as shared tasks (Huang and Lu, 2015). These tasks cover a diverse range of text genres (biomedical literature and clinical notes), dataset sizes, and degrees of difficulty and, more importantly, highlight common biomedicine text-mining challenges. We expect that the models that perform better on all or most tasks in BLUE will address other biomedicine tasks more robustly.

To better understand the challenge posed by BLUE, we conduct experiments with two baselines: One makes use of the BERT model (Devlin et al., 2019) and one makes use of ELMo (Peters et al., 2017). Both are state-of-the-art language representation models and demonstrate promising results in NLP tasks of general purpose. We find that the BERT model pre-trained on PubMed abstracts (Fiorini et al., 2018) and MIMIC-III clinical notes (Johnson et al., 2016) achieves the best results, and is significantly superior to other models in the clinical domain. This demonstrates the importance of pre-training among different text genres.

In summary, we offer: (i) five tasks with ten biomedical and clinical text-mining corpora with different sizes and levels of difficulty, (ii) codes for data construction and model evaluation for fair comparisons, (iii) pretrained BERT models on PubMed abstracts and MIMIC-III, and (iv) baseline results.

2   Related work

There is a long history of using shared language representations to capture text semantics in biomedical text and data mining research. Such re-

58

search utilizes a technique, termed transfer learning, whereby the language representations are pre-trained on large corpora and fine-tuned in a variety of downstream tasks, such as named entity recognition and relation extraction.

One established trend is a form of word embeddings that represent the semantic, using high dimensional vectors (Chiu et al., 2016; Wang et al., 2018c; Zhang et al., 2019). Similar methods also have been derived to improve embeddings of word sequences by introducing sentence embeddings (Chen et al., 2019). They always, however, require complicated neural networks to be effectively used in downstream applications.

Another popular trend, especially in recent years, is the context-dependent representation. Different from word embeddings, it allows the meaning of a word to change according to the context in which it is used (Melamud et al., 2016; Peters et al., 2017; Devlin et al., 2019; Dai et al., 2019). In the scientific domain, Beltagy et al. released SciBERT which is trained on scientific text. In the biomedical domain, BioBERT (Lee et al., 2019) and BioELMo (Jin et al., 2019) were pre-trained and applied to several specific tasks. In the clinical domain, Alsentzer et al. (2019) released a clinical BERT base model trained on the MIMIC-III database. Most of these works, however, were evaluated on either different datasets or the same dataset with slightly different sizes of examples. This makes it challenging to fairly compare various language models.

Based on these reasons, a standard benchmarking is urgently required. Parallel to our work, Lee et al. (2019) introduced three tasks: named entity recognition, relation extraction, and QA, while Jin et al. (2019) introduced NLI in addition to named entity recognition. To this end, we deem that BLUE is different in three ways. First, BLUE is selected to cover a diverse range of text genres, including both biomedical and clinical domains. Second, BLUE goes beyond sentence or sentence pairs by including document classification tasks. Third, BLUE provides a comprehensive suite of codes to reconstruct dataset from scratch without removing any instances.

3 Tasks

BLUE contains five tasks with ten corpora that cover a broad range of data quantities and difficulties (Table 1). Here, we rely on preexisting datasets because they have been widely used by the BioNLP community as shared tasks.

3.1 Sentence similarity

The sentence similarity task is to predict similarity scores based on sentence pairs. Following common practice, we evaluate similarity by using Pearson correlation coefficients.

BIOSSES is a corpus of sentence pairs selected from the Biomedical Summarization Track Training Dataset in the biomedical domain (Soğancıoğlu et al., 2017).[1] To develop BIOSSES, five curators judged their similarity, using scores that ranged from 0 (no relation) to 4 (equivalent). Here, we randomly select 80% for training and 20% for testing because there is no standard splits in the released data.

MedSTS is a corpus of sentence pairs selected from Mayo Clinic's clinical data warehouse (Wang et al., 2018b). To develop MedSTS, two medical experts graded the sentence's semantic similarity scores from 0 to 5 (low to high similarity). We use the standard training and testing sets in the shared task.

3.2 Named entity recognition

The aim of the named entity recognition task is to predict mention spans given in the text (Jurafsky and Martin, 2008). The results are evaluated through a comparison of the set of mention spans annotated within the document with the set of mention spans predicted by the model. We evaluate the results by using the strict version of precision, recall, and F1-score. For disjoint mentions, all spans also must be strictly correct. To construct the dataset, we used spaCy[2] to split the text into a sequence of tokens when the original datasets do not provide such information.

BC5CDR is a collection of 1,500 PubMed titles and abstracts selected from the CTD-Pfizer corpus and was used in the BioCreative V chemical-disease relation task (Li et al., 2016).[3] The diseases and chemicals mentioned in the articles were annotated independently by two human experts with medical training and curation experience. We use the standard training and test set in the

---

[1] http://tabilab.cmpe.boun.edu.tr/BIOSSES/
[2] https://spacy.io/
[3] https://biocreative.bioinformatics.udel.edu/tasks/biocreative-v/track-3-cdr/

Corpus	Train	Dev	Test	Task	Metrics	Domain	Avg sent len
MedSTS, sentence pairs	675	75	318	Sentence similarity	Pearson	Clinical	25.8
BIOSSES, sentence pairs	64	16	20	Sentence similarity	Pearson	Biomedical	22.9
BC5CDR-disease, mentions	4182	4244	4424	NER	F1	Biomedical	22.3
BC5CDR-chemical, mentions	5203	5347	5385	NER	F1	Biomedical	22.3
ShARe/CLEFE, mentions	4628	1075	5195	NER	F1	Clinical	10.6
DDI, relations	2937	1004	979	Relation extraction	micro F1	Biomedical	41.7
ChemProt, relations	4154	2416	3458	Relation extraction	micro F1	Biomedical	34.3
i2b2 2010, relations	3110	11	6293	Relation extraction	F1	Clinical	24.8
HoC, documents	1108	157	315	Document classification	F1	Biomedical	25.3
MedNLI, pairs	11232	1395	1422	Inference	accuracy	Clinical	11.9

Table 1: BLUE tasks

BC5CDR shared task (Wei et al., 2016).

ShARe/CLEF eHealth Task 1 Corpus is a collection of 299 deidentified clinical free-text notes from the MIMIC II database (Suominen et al., 2013).[4] The disorders mentioned in the clinical notes were annotated by two professionally trained annotators, followed by an adjudication step, resulting in high inter-annotator agreement. We use the standard training and test set in the ShARe/CLEF eHealth Tasks 1.

3.3 Relation extraction

The aim of the relation extraction task is to predict relations and their types between the two entities mentioned in the sentences. The relations with types were compared to annotated data. We use the standard micro-average precision, recall, and F1-score metrics.

DDI extraction 2013 corpus is a collection of 792 texts selected from the DrugBank database and other 233 Medline abstracts (Herrero-Zazo et al., 2013).[5] The drug-drug interactions, including both pharmacokinetic and pharmacodynamic interactions, were annotated by two expert pharmacists with a substantial background in pharmacovigilance. In our benchmark, we use 624 train files and 191 test files to evaluate the performance and report the micro-average F1-score of the four DDI types.

ChemProt consists of 1,820 PubMed abstracts with chemical-protein interactions annotated by domain experts and was used in the BioCreative VI text mining chemical-protein interactions shared task (Krallinger et al., 2017).[6] We use the

standard training and test sets in the ChemProt shared task and evaluate the same five classes: CPR:3, CPR:4, CPR:5, CPR:6, and CPR:9.

i2b2 2010 shared task collection consists of 170 documents for training and 256 documents for testing, which is the subset of the original dataset (Uzuner et al., 2011).[7] The dataset was collected from three different hospitals and was annotated by medical practitioners for eight types of relations between problems and treatments.

3.4 Document multilabel classification

The multilabel classification task predicts multiple labels from the texts.

HoC (the Hallmarks of Cancers corpus) consists of 1,580 PubMed abstracts annotated with ten currently known hallmarks of cancer (Baker et al., 2016).[8] Annotation was performed at sentence level by an expert with 15+ years of experience in cancer research. We use 315 (∼20%) abstracts for testing and the remaining abstracts for training. For the HoC task, we followed the common practice and reported the example-based F1-score on the abstract level (Zhang and Zhou, 2014; Du et al., 2019).

3.5 Inference task

The aim of the inference task is to predict whether the premise sentence entails or contradicts the hypothesis sentence. We use the standard overall accuracy to evaluate the performance.

MedNLI is a collection of sentence pairs selected from MIMIC-III (Romanov and Shivade, 2018).[9] Given a premise sentence and a hy-

---

[4] https://physionet.org/works/ShAReCLEFeHealth2013/

[5] http://labda.inf.uc3m.es/ddicorpus

[6] https://biocreative.bioinformatics.udel.edu/news/corpora/

chemprot-corpus-biocreative-vi/

[7] https://www.i2b2.org/NLP/DataSets/

[8] https://www.cl.cam.ac.uk/~sb895/HoC.html

[9] https://physionet.org/physiotools/mimic-code/mednli/

pothesis sentence, two board-certified radiologists graded whether the task predicted whether the premise entails the hypothesis (entailment), contradicts the hypothesis (contradiction), or neither (neutral). We use the same training, development, and test sets in Romanov and Shivade (Romanov and Shivade, 2018).

3.6 Total score

Following the practice in Wang et al. (2018a) and Lee et al. (2019), we use a macro-average of F1-scores and Pearson scores to determine a system's position.

4 Baselines

For baselines, we evaluate several pre-training models as described below. The original code for the baselines is available at `https://github.com/ncbi-nlp/NCBI_BERT`.

4.1 BERT

4.1.1 Pre-training BERT

BERT (Devlin et al., 2019) is a contextualized word representation model that is pre-trained based on a masked language model, using bidirectional Transformers (Vaswani et al., 2017).

In this paper, we pre-trained our own model BERT on PubMed abstracts and clinical notes (MIMIC-III). The statistics of the text corpora on which BERT was pre-trained are shown in Table 2.

Corpus	Words	Domain
PubMed abstract	> 4,000M	Biomedical
MIMIC-III	> 500M	Clinical

Table 2: Corpora

We initialized BERT with pre-trained BERT provided by (Devlin et al., 2019). We then continue to pre-train the model, using the listed corpora.

We released our BERT-Base and BERT-Large models, using the same vocabulary, sequence length, and other configurations provided by Devlin et al. (2019). Both models were trained with 5M steps on the PubMed corpus and 0.2M steps on the MIMIC-III corpus.

4.1.2 Fine-tuning with BERT

BERT is applied to various downstream text-mining tasks while requiring only minimal architecture modification.

For sentence similarity tasks, we packed the sentence pairs together into a single sequence, as suggested in Devlin et al. (2019).

For named entity recognition, we used the BIO tags for each token in the sentence. We considered the tasks similar to machine translation, as predicting the sequence of BIO tags from the input sentence.

We treated the relation extraction task as a sentence classification by replacing two named entity mentions of interest in the sentence with predefined tags (e.g., @GENE$, @DRUG$) (Lee et al., 2019). For example, we used "@CHEMICAL$ protected against the RTI-76-induced inhibition of @GENE$ binding." to replace the original sentence "Citalopram protected against the RTI-76-induced inhibition of SERT binding." in which "citalopram" and "SERT" has a chemical-gene relation.

For multi-label tasks, we fine-tuned the model to predict multi-labels for each sentence in the document. We then combine the labels in one document and compare them with the gold-standard.

Like BERT, we provided sources code for fine-tuning, prediction, and evaluation to make it straightforward to follow those examples to use our BERT pre-trained models for all tasks.

4.2 Fine-tuning with ELMo

We adopted the ELMo model pre-trained on PubMed abstracts (Peters et al., 2017) to accomplish the BLUE tasks.[10] The output of ELMo embeddings of each token is used as input for the fine-tuning model. We retrieved the output states of both layers in ELMo and concatenated them into one vector for each word. We used the maximum sequence length 128 for padding. The learning rate was set to 0.001 with an Adam optimizer. We iterated the training process for 20 epochs with batch size 64 and early stopped if the training loss did not decrease.

For sentence similarity tasks, we used bag of embeddings with the average strategy to transform the sequence of word embeddings into a sentence embedding. Afterward, we concatenated two sentence embeddings and fed them into an architecture with one dense layer to predict the similarity of two sentences.

---

[10] `https://allennlp.org/elmo`

Task	Metrics	SOTA*	ELMo	BioBERT	Our BERT			
					Base (P)	Base (P+M)	Large (P)	Large (P+M)
MedSTS	Pearson	83.6	68.6	84.5	84.5	**84.8**	84.6	83.2
BIOSSES	Pearson	84.8	60.2	82.7	89.3	**91.6**	86.3	75.1
BC5CDR-disease	F	84.1	83.9	85.9	**86.6**	85.4	82.9	83.8
BC5CDR-chemical	F	93.3	91.5	93.0	**93.5**	92.4	91.7	91.1
ShARe/CLEFE	F	70.0	75.6	72.8	75.4	**77.1**	72.7	74.4
DDI	F	72.9	78.9	78.8	78.1	79.4	**79.9**	76.3
ChemProt	F	64.1	66.6	71.3	72.5	69.2	**74.4**	65.1
i2b2	F	73.7	71.2	72.2	74.4	**76.4**	73.3	73.9
HoC	F	81.5	80.0	82.9	85.3	83.1	**87.3**	85.3
MedNLI	acc	73.5	71.4	80.5	82.2	**84.0**	81.5	83.8
Total			78.8	80.5	82.2	**82.3**	81.5	79.2

* SOTA, state-of-the-art as of April 2019, to the best of our knowledge: MedSTS, BIOSSES (Chen et al., 2019); BC5CDR-disease, BC5CDR-chem (Yoon et al., 2018); ShARe/CLEFE (Leaman et al., 2015); DDI (Zhang et al., 2018). Chem-Prot (Peng et al., 2018); i2b2 (Rink et al., 2011); HoC (Du et al., 2019); MedNLI (Romanov and Shivade, 2018). P: PubMed, P+M: PubMed + MIMIC-III

Table 3: Baseline performance on the BLUE task test sets.

For named entity recognition, we used a Bi-LSTM-CRF implementation as a sequence tagger (Huang et al., 2015; Si et al., 2019; Lample et al., 2016). Specifically, we concatenated the GloVe word embeddings (Pennington et al., 2014), character embeddings, and ELMo embeddings of each token and fed the combined vectors into the sequence tagger to predict the label for each token. The GloVe word embeddings[11] and character embeddings have 100 and 25 dimensions, respectively. The hidden sizes of the Bi-LSTM are also set to 100 and 25 for the word and character embeddings, respectively.

For relation extraction and multi-label tasks, we followed the steps in fine-tuning with BERT but used the averaged ELMo embeddings of all words in each sentence as the sentence embedding.

5 Benchmark results and discussion

We pre-trained four BERT models: BERT-Base (P), BERT-Large (P), BERT-Base (P+M), BERT-Large (P+M) on PubMed abstracts only, and the combination of PubMed abstracts and clinical notes, respectively. We present performance on the main benchmark tasks in Table 3. More detailed comparison is shown in the Appendix A.

Overall, our BERT-Base (P+M) that were pre-trained on both PubMed abstract and MIMIC-III achieved the best results across five tasks, even though it is only slightly better than the one pre-trained on PubMed abstracts only. Compared to the tasks in the clinical domain and biomedical domain, BERT-Base (P+M) is significantly superior to other models. This demonstrates the importance of pre-training among different text genres.

When comparing BERT pre-trained using the base settings against that using the large settings, it is a bit surprising that BERT-Base is better than BERT-Large except in relation extraction and document classification tasks. Further analysis shows that, on these tasks, the average length of sentences is longer than those of others (Table 1). In addition, BERT-Large pre-trained on PubMed and MIMIC is worse than other models overall. However, BERT-Large (P) performs the best in the multilabel task, even compared with the feature-based model utilizing enriched ontology (Yan and Wong, 2017). This is partially because the MIMIC-III data are relatively smaller than the PubMed abstracts and, thus, cannot pre-train the large model sufficiently.

In the sentence similarity tasks, BERT-Base (P+M) achieves the best results on both datasets. Because the BIOSSES dataset is very small (there

are only 16 sentence pairs in the test set), all BERT models' performance was unstable. This problem has also been noted in the work of Devlin et al. (2019) when the model was evaluated on the GLUE benchmarking. Here, we obtained the best results by following the same strategy: selecting the best model on the development set after several runs. Other possible ways to overcome this issue include choosing the model with the best performance from multiple runs or averaging results from multiple fine-tuned models.

In the named entity recognition tasks, BERT-Base (P) achieved the best results on two biomedical datasets, whereas BERT-Base (P+M) achieved the best results on the clinical dataset. In all cases, we observed that the winning model obtained higher recall than did the others. Given that we use the pre-defined vocabulary in the original BERT and that this task relies heavily on the tokenization, it is possible that using BERT as pertaining to a custom sentence piece tokenizer may further improve the model's performance.

6 Conclusion

In this study, we introduce BLUE, a collection of resources for evaluating and analyzing biomedical natural language representation models. We find that the BERT models pre-trained on PubMed abstracts and clinical notes see better performance than do most state-of-the-art models. Detailed analysis shows that our benchmarking can be used to evaluate the capacity of the models to understand the biomedicine text and, moreover, to shed light on the future directions for developing biomedicine language representations.

Acknowledgments

References

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew

B. A. McDermott. 2019. Publicly available clinical BERT embeddings. *arXiv:1904.03323*.

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 (Oxford, England)*, 32:432–440.

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

Qingyu Chen, Yifan Peng, and Zhiyong Lu. 2019. BioSentVec: creating sentence embeddings for biomedical texts. In *Proceedings of the 7th IEEE International Conference on Healthcare Informatics*.

Billy Chiu, Gamal Crichton, Anna Korhonen, and Sampo Pyysalo. 2016. How to train good word embeddings for biomedical NLP. In *Proceedings of BioNLP Workshop*, pages 166–174.

Zihang Dai, Zhilin Yang, Yiming Yang, William W Cohen, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jingcheng Du, Qingyu Chen, Yifan Peng, Yang Xiang, Cui Tao, and Zhiyong Lu. 2019. ML-Net: multi-label classification of biomedical texts with deep neural networks. *Journal of the American Medical Informatics Association (JAMIA)*.

Nicolas Fiorini, Robert Leaman, David J Lipman, and Zhiyong Lu. 2018. How user intelligence is improving pubmed. *Nature biotechnology*, 36:937–945.

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*, 46:914–920.

Chung-Chi Huang and Zhiyong Lu. 2015. Community challenges in biomedical text mining over 10 years: success, failure and the future. *Briefings in Bioinformatics*, 17(1):132–144.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Qiao Jin, Bhuwan Dhingra, William W. Cohen, and Xinghua Lu. 2019. Probing biomedical embeddings from language models. *arXiv:1904.02181*.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-Wei H. Lehman, 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.

Daniel Jurafsky and James H. Martin. 2008. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*, 2 edition. Prentice Hall.

Martin Krallinger, Obdulia Rabal, Saber A. Akhondi, Martín Pérez Pérez, Jesús Santamaría, Gael Pérez Rodríguez, Georgios Tsatsaronis, Ander Intxaurrondo, José Antonio López1 Umesh Nandal, Erin Van Buel, Akileshwari Chandrasekhar, Marleen Rodenburg, Astrid Laegreid, Marius Doornenbal, Julen Oyarzabal, Analia Lourenço, and Alfonso Valencia. 2017. Overview of the BioCreative VI chemical-protein interaction track. In *Proceedings of BioCreative*, pages 141–146.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT*, pages 260–270.

Robert Leaman, Ritu Khare, and Zhiyong Lu. 2015. Challenges in clinical natural language processing for automated disorder normalization. *Journal of biomedical informatics*, 57:28–37.

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. *arXiv:1901.08746*.

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.

Oren Melamud, David McClosky, Siddharth Patwardhan, and Mohit Bansal. 2016. The role of context types and dimensionality in learning word embeddings. In *Proceedings of NAACL-HLT*, pages 1030–1040.

Yifan Peng, Anthony Rios, Ramakanth Kavuluru, and Zhiyong Lu. 2018. Extracting chemical-protein relations with ensembles of SVM and deep learning models. *Database: the journal of biological databases and curation*, 2018.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532–1543.

Matthew E. Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of ACL*, pages 1756–1765.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Bryan Rink, Sanda Harabagiu, and Kirk Roberts. 2011. Automatic extraction of relations between medical concepts in clinical texts. *Journal of the American Medical Informatics Association*, 18:594–600.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of EMNLP*, pages 1586–1596.

Yuqi Si, Jingqi Wang, Hua Xu, and Kirk Roberts. 2019. Enhancing clinical concept extraction with contextual embedding. *arXiv preprint arXiv:1902.08691*.

Neil R Smalheiser, Aaron M Cohen, and Gary Bonifield. 2019. Unsupervised low-dimensional vector representations for words, phrases and text that are transparent, scalable, and produce similarity metrics that are not redundant with neural embeddings. *Journal of biomedical informatics*, 90:103096.

Gizem Soğancıoğlu, Hakime Öztürk, and Arzucan Özgür. 2017. BIOSSES: a semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics (Oxford, England)*, 33:i49–i58.

Hanna Suominen, Sanna Salanterä, Sumithra Velupillai, Wendy W Chapman, Guergana Savova, Noemie Elhadad, Sameer Pradhan, Brett R South, Danielle L Mowery, Gareth JF Jones, et al. 2013. Overview of the ShARe/CLEF eHealth evaluation lab 2013. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 212–231. Springer.

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)*, 18:552–556.

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

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018a. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Yanshan Wang, Naveed Afzal, Sunyang Fu, Liwei Wang, Feichen Shen, Majid Rastegar-Mojarad, and Hongfang Liu. 2018b. MedSTS: a resource for clinical semantic textual similarity. *Language Resources and Evaluation*, pages 1–16.

Yanshan Wang, Sijia Liu, Naveed Afzal, Majid Rastegar-Mojarad, Liwei Wang, Feichen Shen, Paul

Kingsbury, and Hongfang Liu. 2018c. A comparison of word embeddings for the biomedical natural language processing. *Journal of biomedical informatics*, 87:12–20.

Chih-Hsuan Wei, Yifan Peng, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Jiao Li, Thomas C Wiegers, and Zhiyong Lu. 2016. Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task. *Database: the journal of biological databases and curation*, 2016.

Shankai Yan and Ka-Chun Wong. 2017. Elucidating high-dimensional cancer hallmark annotation via enriched ontology. *Journal of biomedical informatics*, 73:84–94.

Wonjin Yoon, Chan Ho So, Jinhyuk Lee, and Jaewoo Kang. 2018. CollaboNet: collaboration of deep neural networks for biomedical named entity recognition. *arXiv preprint arXiv:1809.07950*.

Min-Ling Zhang and Zhi-Hua Zhou. 2014. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837.

Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. 2019. BioWordVec, improving biomedical word embeddings with subword information and mesh. *Scientific data*, 6:52.

Yijia Zhang, Wei Zheng, Hongfei Lin, Jian Wang, Zhihao Yang, and Michel Dumontier. 2018. Drug-drug interaction extraction via hierarchical RNNs on sequence and shortest dependency paths. *Bioinformatics (Oxford, England)*, 34:828–835.

A Appendices

TP: true positive, FP: false positive, FN: false negative, P: precision, R: recall, F1: F1-score

A.1 Named Entity Recognition

BC5CDR-disease	TP	FP	FN	P	R	F1
(Yoon et al., 2018)	-	-	-	85.6	82.6	84.1
ELMo	3740	749	684	83.3	84.5	83.9
BioBERT	3807	637	617	85.7	86.1	85.9
Our BERT						
Base (P)	3806	635	564	85.9	87.3	86.6
Base (P+M)	3788	655	636	85.3	85.6	85.4
Large (P)	3729	847	695	81.5	84.3	82.9
Large (P+M)	3765	799	659	82.5	85.1	83.8

BC5CDR-chemical	TP	FP	FN	P	R	F1
(Yoon et al., 2018)	-	-	-	94.3	92.4	93.3
ELMo	4864	386	521	92.6	90.3	91.5
BioBERT	5029	404	356	92.6	93.4	93.0
Our BERT						
Base (P)	5027	336	358	93.7	93.4	93.5
Base (P+M)	4914	341	471	93.5	91.3	92.4
Large (P)	4941	454	444	91.6	91.8	91.7
Large (P+M)	4905	484	480	91.0	91.1	91.1

ShARe/CLEFE	TP	FP	FN	P	R	F1
(Leaman et al., 2015)	-	-	-	79.7	71.3	75.3
ELMo	3928	1117	1423	77.9	73.4	75.6
BioBERT	3898	1024	1453	79.2	72.8	75.9
Our BERT						
Base (P)	4032	1010	1319	80.0	75.4	77.6
Base (P+M)	4126	948	1225	81.3	77.1	79.2
Large (P)	3890	1441	1461	73.0	72.7	72.8
Large (P+M)	3980	1456	1371	73.2	74.4	73.8

A.2 Relation extraction

DDI	TP	FP	FN	P	R	F1
(Zhang et al., 2018)	-	-	-	74.1	71.8	72.9
ELMo	-	-	-	79.0	78.9	78.9
BioBERT	786	229	193	77.4	80.3	78.8
Our BERT						
Base (P)	737	172	242	81.1	75.3	78.1
Base (P+M)	775	198	204	79.7	79.2	79.4
Large (P)	788	206	191	79.3	80.5	79.9
Large (P+M)	748	234	231	76.2	76.4	76.3

Chem-Prot	TP	FP	FN	P	R	F1
(Peng et al., 2018)	1983	746	1475	72.7	57.4	64.1
ELMo	-	-	-	66.7	66.6	66.6
BioBERT	2359	803	1099	74.6	68.2	71.3
Our BERT						
Base (P)	2443	834	1015	74.5	70.6	72.5
Base (P+M)	2354	996	1104	70.3	68.1	69.2
Large (P)	2610	948	848	73.4	75.5	74.4
Large (P+M)	2355	1423	1103	62.3	68.1	65.1

i2b2	TP	FP	FN	P	R	F1
(Rink et al., 2011)	-	-	-	72.0	75.3	73.7
ELMo	-	-	-	71.2	71.1	71.1
BioBERT	4391	1474	1902	74.9	69.8	72.2
Our BERT						
Base (P)	4592	1459	1701	75.9	73.0	74.4
Base (P+M)	4683	1291	1610	78.4	74.4	76.4
Large (P)	4684	1805	1609	72.2	74.4	73.3
Large (P+M)	4700	1719	1593	73.2	74.7	73.9

A.3 Document classification

HoC	P	R	F1
(Du et al., 2019)	81.3	81.7	81.5
ELMo	78.2	81.9	80.0
BioBERT	83.4	82.4	82.9
Our BERT			
Base (P)	86.2	84.4	85.3
Base (P+M)	84.0	82.3	83.1
Large (P)	91.0	83.9	87.3
Large (P+M)	88.8	82.1	85.3