Improving Biomedical Named Entity Recognition using custom NER models

Christine Bakan, Emerald Swei, Viola Pu
UC Berkeley
{christine.bakan, emerald.swei, viola.pu}@berkeley.edu

December 3, 2022

Abstract

Accurate detection of biomedical domain specific entities in scientific literature can lead to faster knowledge discovery in the rapidly growing biomedical research space. This paper presents approaches for custom biomedical NER models fine tuned on three corpora: literature abstracts from the National Center for Biotechnology Information (NCBI) database, annotated sentences from the BioCreative II Gene Mention Recognition Dataset (BC2GM, or GM), and articles from the BioCreative V Chemical Disease Relation (BC5CDR, or CDR) corpus. We will test a total of 6 model variations on BERT and BioBERT, a version of BERT that has been pre-trained on biomedical specific data. We have found that when not combined with downstream layers, BioBERT consistently outperformed BERT. Furthermore, adding BiLSTM and CRF layers in general improve the model performance, but the number of LSTM units need to be fine-tuned depending on the dataset. We also demonstrated transfer learning with an improvement of model performance when trained on additional novel data.

1. Introduction

Named Entity Recognition (NER) is a fundamental task in information extraction, as well as one of the most important tasks in the field of biomedical text mining. With the constant creation and development of new biomedical literature, advanced text mining and entity recognition techniques specific to the biomedical domain are tasked to extract relevant information from the growing corpora of literature.

The ability to detect mentions of biologically relevant entities such as disease, genes, chemicals and drug names in biomedical texts is further complicated as biological entities (i) continually increase with new discoveries, (ii) have growing numbers of synonyms, (iii) are often referred to using abbreviations, (iv) are described by long phrases, and (v) are mixtures of letters, symbols and punctuations (Cho et al, 2019). In recent years, various approaches like supervised machine learning to deep learning methods have shown significant improvement over early rule-based methods for domain entity-specific NER tasks (Song et al, 2021). With the introduction of a new deep learning architecture named BERT (Devlin et al, 2019), new state-of-the-art results on NLP tasks, including NER, using pre-trained self-attention based transformer architectures have been established on non-domain specific corpora of text. Subsequently, BioBERT (Lee et al, 2019), a descendant of BERT that is pre-trained on biomedical corpora, demonstrated improved performance over BERT on biomedical entity recognition tasks.

In this paper, we propose an experiment on a variation of neural network based models for biomedical entity recognition that includes pre-trained BERT and BioBERT models with added bidirectional LSTM

layers (BILSTM) and Conditional Random Field (CRF) layers (BERT+CRF, BioBERT+CRF, BERT+BiLSTM+CRF and BioBERT+BiLSTM+CRF) to demonstrate improved performance of the BiLSTM+CRF model trained using BERT and BioBERT embeddings, over unlayered BERT and BioBERT models. In addition, we also attempt to demonstrate transfer learning with an improvement of model performance when trained on additional novel data. Following the common practice for evaluating NER tasks, we will evaluate the model performance using precision, recall and F-1 score.

1.1 Prior Work

Researchers have applied deep learning techniques for biomedical entity recognition on various medical text corpora.

Li et al compared several neural models, including BiLSTM+CRF, BERT and BERT+BiLSTM+CRF on Chinese clinical texts obtained through web crawling and demonstrated that the fine-tuned BERT+BiLSTM+CRF outperforms both BiLSTM+CRF and BERT base models.

Zhou et al proposed a NER framework using BiLSTM+CRF and BioBERT+CRF with and without a custom labeling for the NCBI Disease corpus and BioCreative V chemical-disease relation (BC5CDR) entities from PubMed. Their findings showed that the baseline BioBERT+CRF model trained on the dataset without any label correction performed better than the BiLSTM+CRF model. Additionally, the fine-tuned BioBERT+CRF model trained on the custom labeled datasets further outperformed the baseline BioBERT+CRF model.

Chen et al experimented with six pre-trained models with a classification layer (using softmax activation) and BiLSTM+CRF as the NER task layers to follow the pre-trained BERT*base, uncased* and BioBERT models on the in-house clinical trial datasets from Covance/LabCorp. Their findings showed that BioBERT+BiLSTM+CRF produced the best precision and F1-scores.

Sun et al demonstrated transfer learning with multilingual pre-trained BERT and English biomedical text pre-trained BioBERT models and fine-tuned on Spanish PharCoNER corpus of chemical and protein entities, where both BERT and BioBERT demonstrated comparable high performance.

2. Methods

This section introduces our datasets and the models we used, including the baseline BERT and BioBERT models as well as the custom models with additional layers on top of BERT and BioBERT.

2.1 Dataset

We used three expert annotated biomedical corpora from NCBI and BioCreative databases representing diseases, genes, proteins, disease-chemical relations, available from Hugging Face in our experiments (Table 1).

In addition, we leveraged a private PubMed literature dataset labeled by 3rd-party biomedical experts (Strand Life Sciences) for transfer learning. This private dataset was provided in .txt and .ann format, and the Disease, Gene-Protein and Chemical entities were labeled at each letter position. We parsed the data into .json format that is similar to the NCBI, GM and CDR datasets, and created labels at word level. We

further separated the 661 abstracts into 7074 records at sentence level, to avoid feeding the models data that is too long to consume.

Dataset	Entity Type	# of Annotations	# of Test / Validation / Train Data
NCBI Disease (Dogan et al, 2014)	Disease	6,881	5433 / 924 / 941
BioCreative II Gene Mention (BC2GM) (Smith et al., 2008)	Gene/ Protein	20,703	12501 / 2501 / 5001
BioCreative V Chemical Disease Relation (BC5CDR) (Li et al., 2016)	Chemicals, Diseases, Chemical-Disease Relations	28,797	4561 / 4582 / 4798
Biomedical expert-labeled PubMed dataset	Disease, Gene/Protein, Chemicals	32, 831	4951 / 1062 / 1062

Table 1: Biomedical named entity recognition datasets

2.1 Modeling approach

To establish a baseline, we use the BERT*large-NER* and BioBERT implementations from Hugging Face, fine-tune for 5 epochs and with a batch size of 8, using the Adam optimizer at a learning rate of 5e-5. We use a softmax-activated Dense layer for the outputs to create a probability distribution for each of the predicted label classes. (Figure 1)

Our custom model architecture (Figure 1) uses the activations from the BERT*large-NER* and BioBERT models as inputs to the downstream layers. A single NCBI Disease corpus example output from BERT/BioBERT would have a shape of (180, 1024)/(180, 768) respectively, where 180 is the max length of the input, and 1024/768 is the dimension of the BERT/BioBERT embeddings. These embeddings are then passed to the downstream layers (CRF or BiLSTM+CRF).

Bidirectional LSTM (BiLSTM) enables sequential dependency modeling through forward and backward sequencing of words. In our experiments, the BiLSTM outputs are fed into the CRF layer. While BERT/BioBERT itself is a bidirectional model, we look to determine whether an introduction of the BiLSTM layer, despite the increased training time, improves performance on NER tasks.

The Conditional Random Fields (CRF) is often used as a post-processing step to refine the segmentation output of the training layer (Lample et al). In our architecture, CRF takes into account the neighboring labels (i.e. the context of the word embeddings) generated as outputs from the BERT, BioBERT or the BiLSTM layer, and produces the final prediction and output of the label.

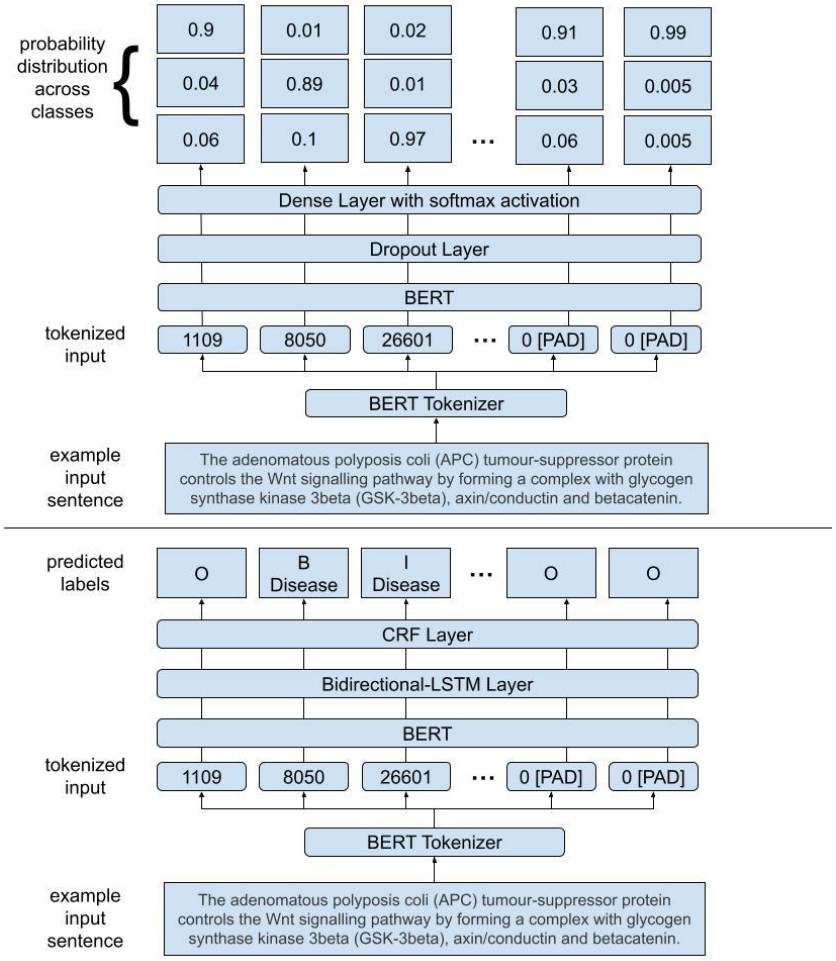

Figure 1: Custom Model Architecture (BERT above, BERT+BiLSTM+CRF below)

We chose to use BERT*large-NER* as our BERT base model for experimentation against BioBERT*cased-v1.1* as our BioBERT base model, based on the NER performance of BERT*large-NER* and four other versions of BERT, against one of the datasets, the NCBI Disease corpus. BERT*large-NER* F-1 score (0.8637) outperformed the other BERT models, making it comparable to BioBERT (Table A1). However, given that there is almost a 3x difference in the # of parameters for BERT*large-NER* vs. BioBERT*cased-v1.1*, we note the trade-off of the higher computational training time for larger models.

3. Results & Interpretation

Fine-tuning custom models on NCBI Disease, BC2GM and BC5CDR datasets

3.1. BERT vs. BioBERT

To establish a baseline, we fine-tuned BERT*large-NER* and BioBERT*cased-v1.1* for 5 epochs with a subset of 100 examples of the NCBI Disease, BC2GM and BC5CDR datasets, followed by 3 epochs with the full datasets.

Table A1 shows that across the NCBI, GM and CDR dataset, when not combined with downstream layers, BioBERT*cased-v1.1* consistently outperformed BERT*large-NER* in precision, recall and F-1 score, even though the latter has 3x more parameters (108,312,579 vs. 333,582,339). This finding corroborates prior literature that BioBERT, being pre-trained on biomedical corpora, demonstrates improved performance over BERT on biomedical entity recognition tasks (Lee et al, 2019), and additionally demonstrates that BioBERT can perform even against a higher-complexity BERT*large-NER*.

3.2. Adding BiLSTM and CRF layers in general improve the model performance, but the number of BiLSTM units need to be fine-tuned depending on the dataset

Table A1 further shows that adding the CRF layer on top of BERT/BioBERT consistently improves the performance of the baseline models across the 3 datasets in most of our chosen metrics of precision, recall and F-1. By comparing results between adding the CRF layer and adding the BiLSTM+CRF layers, we found that the latter only outperforms the former when the number of LSTM units are fine-tuned for the specific dataset and shape of upstream embeddings. We tested different numbers of LSTM units for our BERT/BioBERT+BiLSTM+CRF models, particularly for the NCBI dataset. Among the numbers of LSTM units we tested (16, 32, 64, 512, 680, 768 and 1024), we found that the BERT+BiLSTM+CRF model only outperformed the BERT+CRF model when the number of LSTM units was 16, while the BioBERT+BiLSTM+CRF model only outperformed the BioBERT+CRF model when the number of LSTM units was 680. This could potentially be due to the fact that the embedding sizes of BERT*large-NER* and BioBERT*cased-v1.1* are different (1024 vs. 768), thus requiring different shapes for the downstream LSTM cells to render the best results.

3.3. The best-performing model is inconsistent across datasets

In our experiment, the best-performing model architectures vary across the 3 datasets. For the NCBI dataset, the BERT+BiLSTM+CRF model achieved overall best performance with Precision=0.8741, Recall=0.8851 and F1=0.8796. For the BC2GM dataset, BERT+CRF achieved the best performance with Precision=0.8535, Recall=0.8644 and F1=0.8589. Lastly, for the BC5CDR dataset, BioBERT+CRF achieved best performance with Precision=0.8963, Recall=0.9114 and F1=0.9038. We believe that this could potentially be due to the following reasons: 1) the token lengths are different for each dataset. For NCBI, the longest example has 180 tokens, while that number for GM and CDR are approximately 300. Different input sizes might work best with different model structures and can impact how quickly the model learns. 2) if we had the chance to fine-tune the LSTM units for GM and CDR models, the results could have been different. Due to resource constraints, we were unable to experiment as thoroughly as we would like to with those model combinations and as such, the LSTM layers for these datasets may not be optimized yet.

Transfer learning w/ Custom annotated PubMed data

For the transfer learning experiment, our modeling approach is as follows: we establish the baseline by training BioBERT*cased-v1.1* on NCBI training dataset for 5 epochs, then evaluate the performance on NCBI test dataset. For the treatment, we first train BioBERT*cased-v1.1* on the private expert-labeled PubMed dataset, then further train this model on NCBI training dataset, each for 5 epochs, and finally evaluate the performance on the NCBI test dataset.

3.4 Value-add of transfer learning

Our results demonstrate that the transfer learning approach outperforms the baseline models that were trained and tested on NCBI dataset alone (Table A.3). Although the PubMed dataset contained additional entities to diseases, the model trained on both the PubMed and NCBI dataset was able to capture more disease entities with better precision and recall than the model trained on NCBI alone. This demonstrates the adaptability of our model and the ability to hone in on a very specific task: identifying diseases in text.

4. Conclusion

As a conclusion, we have shown that additional custom layering on baseline models that were pre-trained on general text corpora, produces improved domain specific NER results when compared to the performance of baseline models alone. Based on the experiments with BiLSTM+CRF layers on top of BERT and BioBERT, the tuning of the LSTM units parameter has a tangible impact on the performance of the custom layered models (i.e. BERT+BiLSTM+CRF as well as BioBERT+BiLSTM+CRF). We also believe that the number of label classes and shape of input examples (i.e. the maximum length of the token inputs) has a relationship with the number of LSTM cells needed to optimize model performance. These factors add to the complexity of the data, which can be mitigated with a larger number of LSTM cells, but there is a tradeoff to increasing the number of cells as well. We also found that higher number of epochs invariably increased the performance of our models, but due to constraints on time and computational power, we limited our training to 3-5 epochs depending on the size of the dataset.

References

Cho, H., Lee, H. Biomedical named entity recognition using deep neural networks with contextual information. *BMC Bioinformatics* 20, 735 (2019). https://doi.org/10.1186/s12859-019-3321-4

Song, Bosheng, et al., Deep learning methods for biomedical named entity recognition: a survey and qualitative comparison, *Briefings in Bioinformatics* 22.6 (2021): bbab282

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.

Jinhuk Lee, et al., BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics (2019). doi: 10.1093/bioinformatics/btz682

Xiangyang Li, Huan Zhang, Xiao-Hua Zhou, Chinese clinical named entity recognition with variant neural structures based on BERT methods, Journal of Biomedical Informatics, Volume 107, 2020, 103422, ISSN 1532-0464, https://doi.org/10.1016/j.jbi.2020.103422.

Zhou, H., Liu, Z., Lang, C. *et al.* Improving the recall of biomedical named entity recognition with label re-correction and knowledge distillation. *BMC Bioinformatics* 22, 295 (2021). https://doi.org/10.1186/s12859-021-04200-w

Chen, Miao et al. "Using Pre-trained Transformer Deep Learning Models to Identify Named Entities and Syntactic Relations for Clinical Protocol Analysis." *AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering* (2020).

Cong Sun and Zhihao Yang. 2019. Transfer Learning in Biomedical Named Entity Recognition: An Evaluation of BERT in the PharmaCoNER task. In *Proceedings of the 5th Workshop on BioNLP Open Shared Tasks*, pages 100–104, Hong Kong, China. Association for Computational Linguistics.

Lample, Guillaume, et al. "Neural architectures for named entity recognition." *arXiv preprint arXiv:1603.01360* (2016).

Appendices

A.1 BERT and BioBERT fine-tuning results

BERT*large-NER and* BioBERT*cased-v1.1* were fine tuned for the NCBI, BC2GM and BC5CDR datasets.

Dataset	Model version	BiLSTM units	# of epochs (sample/full dataset)	Precision	Recall	F1
NCBI Disease	BERT	-	5/5	0.8692	0.8833	0.8762
	BERT+CRF	-	5/5	0.8678	0.8577	0.8627
	BERT+BiLSTM+CRF	16	5/5	**0.8741**	0.8851	**0.8796**
	BERT+BiLSTM+CRF	32	5/5	0.8640	0.8811	0.8724
	BERT+BiLSTM+CRF	64	5/5	0.8648	0.8613	0.8630
	BERT+BiLSTM+CRF	768	5/5	0.8049	**0.9032**	0.8512
	BERT+BiLSTM+CRF	1024	5/5	0.8575	0.8838	0.8705
	BioBERT	-	5/5	0.8607	0.8697	0.8654
	BioBERT+CRF	-	5/5	0.8539	0.8714	0.8626
	BioBERT+BiLSTM+CRF	16	5/5	0.8428	0.8320	0.8625
	BioBERT+BiLSTM+CRF	512	5/5	0.8537	**0.8954**	0.8740
	BioBERT+BiLSTM+CRF	680	5/5	**0.8707**	0.8862	**0.8784**
	BioBERT+BiLSTM+CRF	1024	5/5	0.8393	0.8697	0.8542
BC2GM	BERT		5/3	0.8269	0.8565	0.8415
	BERT+CRF		5/3	**0.8535**	0.8644	**0.8589**
	BERT+BiLSTM+CRF	768	5/3	0.8129	**0.8815**	0.8458
	BioBERT		5/3	0.8364	**0.8724**	**0.8540**
	BioBERT+CRF		5/3	**0.8376**	0.8647	0.8510
	BioBERT+BiLSTM+CRF	768	5/3	0.8128	0.8247	0.8187
BC5CDR	BERT		5/5	0.8804	0.9041	0.8918
	BERT+CRF		5/5	**0.8855**	**0.9042**	**0.8947**

BERT+BiLSTM+CRF	768	5/5	0.8812	0.8961	0.8885
BioBERT		5/5	0.8766	**0.9176**	0.8965
BioBERT+CRF		5/5	0.8963	0.9114	**0.9038**
BioBERT+BiLSTM+CRF	768	5/5	**0.9000**	0.8860	0.8929

A.2 BERT and BioBERT performance evaluation against NCBI Disease corpus

Model version	# of Parameters	# of epochs (sample/full dataset)	Precision	Recall	F1
BERTbase-cased	108,312,579	5/3	0.8266	0.8590	0.8425
BERT*base-uncased*	109,484,547	5/3	0.8592	0.8488	0.8539
BERT*large-cased*	333,582,339	5/3	**0.8606**	0.8482	0.8544
BERT*large-uncased*	335,144,963	5/3	0.8212	0.8874	0.8530
BERT*large-NER*	333,582,339	5/3	0.8500	**0.8779**	**0.8637**
BioBERT*cased-v1.1*	108,310,272	5/3	0.8395	**0.8963**	**0.8670**

A.3 Transfer Learning

Model version	Dataset	# of epochs (full dataset)	Precision	Recall	F1
BioBERT*cased-v1.1*	NCBI	5 (NCBI)	0.8076	0.9002	0.8514
BioBERT*cased-v1.1*	Abstract + NCBI	5 (Abstract 5 (NCBI)	**0.8379**	**0.8921**	**0.8642**