Transfer Learning for Biomedical Named Entity Recognition with BioBERT

Anthi Symeonidou[1,2], Viachaslau Sazonau[2], and Paul Groth[1]

[1] University of Amsterdam, Amsterdam, The Netherlands
anthi.symeonidou@student.uva.nl, p.groth@uva.nl
[2] Elsevier, Amsterdam, The Netherlands
s.sazonau@elsevier.com

Abstract. We apply a transfer learning approach to biomedical named entity recognition and compare it with traditional approaches (dictionary, CRF, BiLTSM). Specifically, we build models for adverse drug reaction recognition on three datasets. We tune a pre-trained transformer model, BioBERT, on these datasets and observe the absolute F1-score improvements of 6.93, 10.46 and 13.31. This shows that, with a relatively small amount of annotated data, transfer learning can help in specialized information extraction tasks.

Keywords: Named entity recognition · BioBERT · Transfer learning · Text mining · BIO tagging · Drug safety · Adverse drug reaction.

1 Introduction

Biomedical knowledge graphs are becoming important for tasks such as pharmacovigilance [14]. These knowledge graphs are often constructed by performing information extraction (IE) over unstructured and semi-structured sources such as clinical records, electronic health records and biomedical literature [2, 9, 13]. Named entity recognition (NER) is one of the fundamental tasks of IE. Particular entities of interest in this domain are adverse drug reactions (ADRs). ADRs cause significant number of deaths worldwide and billion of dollars are spent yearly to treat people who had an ADR from a prescribed drug [11]. ADR recognition is a challenging IE task because of the context importance and multi-token phenomena such as discontinuity and overlaps.

Various approaches have been applied to biomedical NER. Dictionary-based approaches [15], where string matching methods are used to identify entities in text, are common [4]. Recently, machine learning techniques, such as conditional random fields (CRFs) and deep learning [3, 6, 10], have gained popularity and shown performance gains. However, these techniques usually require lots of training data, which is costly to obtain in the biomedical domain.

Transfer learning techniques have shown their potential in overcoming the lack of training data. Transfer learning is a method where a model developed for one task is exploited to improve generalization on another task [16]. Giorgi and Bader have found that a transfer learning approach is beneficial for biomedical NER and can improve state-of-the-art results, particularly for datasets with a number of labels less than 6000 [7]. Devlin et al, have recently published a model called BERT (Bidirectional Encoder Representations from Transformers), which was trained over 3.3B words corpus and achieved outstanding performance in 11 natural language processing tasks (NLP), including NER [5]. In the biomedical context, BioBERT, which has a similar architecture to BERT and was trained on more than 18B words from PubMed abstracts and PMC articles, achieved high performance in NER on several benchmarks [12].

In this paper, we investigate if transfer learning can outperform traditional approaches for ADR recognition on three different datasets. The pre-trained BioBERT model was fine-tuned on these datasets and compared to a dictionary-based method, CRF, and BiLTSM. Our main contribution is empirical and shows that a transfer learning method based on BioBERT can achieve considerably higher performance in recognizing ADRs than traditional methods. Our results suggest that transfer learning based on transformer architectures is a promising approach to addressing the lack of training data in biomedical IE.

2 Data & Models

In our experiments, we use two open benchmarks for ADR recognition (the only ones publicly available for ADR recognition to the best of our knowledge) and complement them with Elsevier's dataset for better reliability of the results.

TAC2017: The data are spontaneous adverse event reports submitted to the FDA Adverse Event Reporting System (FAERS) by drug manufacturers in a Structured Product Labeling (SPL) format from 2009. These `xml` documents are converted to Brat standoff format [18].

ADE corpus: ADE is an open source corpus which consists of information extracted by PubMed articles and contains annotated entities (ADRs, drugs, doses) and relations. The ADRs annotations were used in the current research [8].

Elsevier's gold set: Elsevier's gold set consists of `xml` text files which come from DAILYMED in SPL format and contain information for human drugs from 2015 to present [3]. They are similar to TAC2017 SLP documents and follow the same annotation guidelines.

Three traditional NER approaches were compared with transfer learning.

Dictionary-Based Approach. A dictionary-based approach is based on the Aho-Corasick algorithm which is a string-matching algorithm. The output is a dictionary of keywords for a given entity type used to create a finite state machine for searching [1].

CRFs. Conditional Random Fields is a probabilistic graphical model that predicts a sequence of labels for a sequence of input samples (sentences in our case). The output is a probability between 0 and 1 which denotes how well the model predicts and assigns the correct label to entities based on features and weights [3].

BiLSTM-CRF. We use a bidirectional LSTM model with a sequential conditional random layer above it. Character and word embeddings were given as input to the model [10]. The best F_1-score reported for TAC2017 was achieved

[3] https://dailymed.nlm.nih.gov/dailymed/

by using the same architecture [18].

BioBERT. A transformer-based model which is based on BERT but pre-trained in a bidirectional way over large-scale biomedical corpora. This model, with minimal task-specific modifications (fine-tuning), was applied to the biomedical NER task [12].

3 Experiments

In our evaluation,[4] we focus on predicting the ADR label as it is of great importance and is one of the predominant labels, representing 9-14% of the total number of words in all our datasets. Due to computational constraints, only sentences with less than 130 words were selected. This filtered out between 0-9% of the sentences depending on the dataset. Duplicate sentences were removed. The number of sentences, words and labelled entities of all datasets are shown in Table 1. The same data pre-processing was accomplished for all approaches and all datasets. The NLTK tokenizer[5] was used for the conventional models. For BioBERT, we used the default WordPiece tokenizer. In order to minimise tokenisation differences, we implemented post-processing of its output to make it similar to the output of the NLTK tokenizer. We use 5-fold cross validation to split the processed data and evaluate the models. In the BioBERT case, the model was fine-tuned on an NVIDIA Tesla T4 16GB GPU using the default hyperparameter values and set only the maximum sequence length to 150 tokens. The official BioBERT implementation was used.[6] Entities are marked based on the standard BIO tagging scheme. The standard precision, recall and F1-score were used as entity-level evaluation metrics where only exact matches of the full entity are counted (i.e. both "B" and "I" tags must match if they belong to the same entity) and "O" is excluded.

Table 1. Biomedical Datasets

	TAC2017	ADE	Elsevier's gold set
Number of sentences	7.934	4.271	3.794
Number of sentences <130 words	7.408	4.271	3.675
Number of Words	207.241	85.832	99.054
Number of ADR labels	16.469	12.226	7.123

4 Results

The mean scores and 95% confidence intervals are shown in Table 2. BioBERT outperformed all other models on all datasets. The absolute improvement compared to the second-best results was 6.93, 10.46 and 13.31 units on TAC2017, ADE and Elsevier's gold set, respectively.[7] Interestingly, the results show that BioBERT achieved much higher recall improvements than precision improvements. This means that the model was able to miss much fewer ADR entities

[4] Source code: https://github.com/AnthiS/MasterThesis_DS

[5] https://www.nltk.org

[6] https://github.com/dmis-lab/biobert

[7] The best F1-score reported on TAC2017 is 85.2% [18] and 86.78% on ADE [17].

Table 2. Models Performance in ADR Named Entity Recognition. Precision (P), Recall (R) and F1-score (F). Best scores are in bold, while second best are underlined.

Dataset	Metric	Dictionary	CRFs	BiLSTM	BioBERT
TAC2017	P	65.57 (±1.38)	83.62 (±1.77)	85.84 (±1.27)	90.90 (±0.97)
	R	82.89 (±1.25)	80.04 (±1.18)	85.13 (±1.59)	93.98 (±0.91)
	F	73.20 (±0.69)	81.77 (±0.85)	85.47 (±0.73)	**92.40** (±0.67)
ADE	P	59.02 (±1.05)	74.50 (±0.83)	72.75 (±2.86)	82.00 (±0.95)
	R	58.80 (±0.95)	69.77 (±1.21)	74.65 (±1.83)	86.33 (±0.70)
	F	58.91 (±0.99)	72.05 (±0.91)	73.65 (±1.81)	**84.11** (±0.78)
Elsevier's gold set	P	64.87 (±2.25)	80.25 (±1.12)	75.42 (±3.20)	86.15 (±2.09)
	R	63.19 (±2.71)	69.38 (±3.09)	72.48 (±2.20)	89.27 (±0.97)
	F	63.96 (±1.53)	74.39 (±1.94)	73.84 (±1.20)	**87.70** (±1.39)

than the other methods. The simultaneous bidirectional contextual information capturing, an important characteristic of BioBERT, seems to be beneficial for model performance. Another important remark is about the number of annotated data that have been used and the corresponding model performance. In particular, only around 7000 of labelled examples were needed in the case of the Elsevier's gold set to achieve the results above 80% F1-score. No significant differences were observed for training time between the BiLSTM and BioBERT model, considering the boost in performance. The fine-tuning of BioBERT requires about 30 minutes, while the BiLSTM needs about 20 minutes on the TAC2017 dataset. However, a GPU is required for the pre-trained BioBERT model due to its high number (110M) of parameters. Overall, the BioBERT-based approach clearly outperformed the traditional approaches for ADR recognition on all our datasets.

5 Conclusion

A transfer learning approach for ADR recognition was tested using a domain specific language model - BioBERT. The fine-tuned BioBERT model achieved better performance on three different biomedical corpora than three traditional methods. An interesting observed property of this model is its ability to find more entities compared to existing methods, using only a few thousand examples and requiring comparable amounts of training time. Based on our results, we believe that transformer-based neural models are a promising approach for complex biomedical NER problems, such as ADR recognition, and can be a key ingredient for biomedical knowledge graph construction. Additional experiments on more datasets, for ADR and other entities, should bring more empirical evidence to validate our conjecture.

References

1. Aho, A.V., Corasick, M.J.: Efficient string matching: An aid to bibliographic search. Commun. ACM **18**(6), 333–340 (1975)

2. Aramaki, E., Miura, Y., Tonoike, M., Ohkuma, T., Masuichi, H., Waki, K., Ohe, K.: Extraction of adverse drug effects from clinical records. Studies in health technology and informatics **160**, 739–43 (2010)
3. Bundschus, M., Dejori, M., Stetter, M., Tresp, V., Kriegel, H.P.: Extraction of semantic biomedical relations from text using conditional random fields. BMC Bioinformatics (2008)
4. Cohen, K.B., Hunter, L.: Getting started in text mining. PLOS Computational Biology **4**(1), 1–3 (2008)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. CoRR **abs/1810.04805** (2018)
6. Dong, X., Qian, L., Guan, Y., Huang, L., Yu, Q., Yang, J.: A multiclass classification method based on deep learning for named entity recognition in electronic medical records. In: 2016 New York Scientific Data Summit (NYSDS). pp. 1–10 (2016)
7. Giorgi, J., D Bader, G.: Transfer learning for biomedical named entity recognition with neural networks. Bioinformatics (Oxford, England) **34** (2018)
8. Gurulingappa, H., Rajput, A.M., Roberts, A., Fluck, J., Hofmann-Apitius, M., Toldo, L.: Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. Journal of Biomedical Informatics **45**(5), 885 – 892 (2012)
9. Harpaz, R., Vilar, S., DuMouchel, W., Salmasian, H., Haerian, K., Shah, N.H., Chase, H.S., Friedman, C.: Combing signals from spontaneous reports and electronic health records for detection of adverse drug reactions. Journal of the American Medical Informatics Association **20**(3), 413–419 (2012)
10. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 260–270. Association for Computational Linguistics, San Diego, California (2016)
11. Lazarou, J., Pomeranz, B.H., Corey, P.N.: Incidence of Adverse Drug Reactions in Hospitalized PatientsA Meta-analysis of Prospective Studies. JAMA **279**(15), 1200–1205 (1998)
12. Lee, J., Yoon, W., Kim, S., Donghyeon, K., Kim, S., Ho So, C., Kang, J.: Biobert: pre-trained biomedical language representation model for biomedical text mining (2019)
13. Leser, U., Hakenberg, J.: What makes a gene name? Named entity recognition in the biomedical literature. Briefings in Bioinformatics **6**(4), 357–369 (2005)
14. Lu, Z.: Pubmed and beyond: a survey of web tools for searching biomedical literature. Database : the journal of biological databases and curation **2011**, baq036 (2011)
15. Ono, T., Hishigaki, H., Tanigami, A., Takagi, T.: Automated extraction of information on protein-protein interactions from the biological literature. Bioinformatics **17**(2), 155–161 (2001)
16. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering **22**(10), 1345–1359 (2010)
17. Ramamoorthy, S., Murugan, S.: An attentive sequence model for adverse drug event extraction from biomedical text (2018)
18. Roberts, K., Demner-Fushman, D., Tonning, J.M.: Overview of the tac 2017 adverse reaction extraction from drug labels track. In: TAC (2017)