Named Entity Recognition for Adverse Drug Events: A Comparative Study on the CADEC Corpus

Mulakkayala Sai Krishna Reddy s4238206
s.k.r.mulakkayala@umail.leidenuniv.nl
LIACS, Leiden University
Leiden,Netherlands

Uzer Ahmed s4506324
u.ahmed@umail.leidenuniv.nl
LIACS, Leiden University
Leiden,Netherlands

Abstract

The identification of adverse drug events (ADEs) is an important concern in healthcare as they hide in plain sight as text data located in clinical notes or online discussion boards. NER as one of the basic data extraction NLP tasks helps to identify entities and produce pharmacovigilance and other clinically relevant information. In this work, we investigate NER on the CSIRO Adverse Drug Event Corpus (CADEC), evaluating the classical approaches as CRF against the more recent transformer-based deep learning models as BioBERT. In this paper, we evaluate the effects of pre-processing and domain-specific embeddings on the performance of the models. Observations show that popularity-based embeddings greatly improve the recognition accuracy of ADE-related entities within the domains. The results presented in this paper are beneficial for creating effective and highly accurate NER systems by addressing the medical domain.

Keywords

ADEs, NER, CADEC Corpus, BioBERT, NLP, Healthcare

1 Introduction

ADEs are a major problem in the healthcare setting causing increased morbidity, mortality and healthcare costs globally. This paper aims to discuss some necessary steps in the mild and reporting of: adverse drug events for the sake of enhancing the safety of patients and the efficacy of the medicine. However, with ADEs the relevant information can be found in the unstructured context of medical text, including social media groups, clinical notes, and articles. It can therefore be noted that extracting such information is quite a problem for natural language processing (NLP). Named Entity Recognition or NER is a prerequisite task in NLP focusing on the identification of difficult items like drugs or symptoms, and diseases. Concerning ADE detection, NER systems can contribute structured outputs that can increase utility within pharmacovigilance activities and improve clinical decision-making. The case study of using the CSIRO Adverse Drug Event Corpus (CADEC) points out the usefulness of this resource for training and testing of NER models for the medical field. Housing entities like drugs, diseases, and symptoms, CADEC corpus has its troubles like terminology peculiar to pharmaceuticals, use of abbreviations and semiotic expressions. This study analyzes the results of traditional and transformer models as applied to the CADEC dataset. Specifically, we address the following research questions:

1. Based on the CADEC dataset, there is a comparison between NER models such as CRF and transformer base models such as BioBERT.

2. This research question concerns the effects of pre-processing steps including lemmatization and tokenization on NER in the context of the medical domain.

3. Can the use of mapping of domain-related embeddings enhance the identification of ADE-related entities?

Hence, answering these questions, we attempt to enrich the understanding of how appropriate NER designs can be employed for medical text as well as unveiling peculiarities of using the approach within the sphere.

2 Related work

In general, Named Entity Recognition (NER) recognizes and categorizes drugs and diseases in a text but specific tasks like adverse drug event (ADE) detection fail sometimes due to the presence of specific terms and synonyms or abbreviations. The CSIRO Adverse Drug Event Corpus (CADEC) was developed by Karimi et al. (2015) and involves 1289 human-posted forums with annotated entities including drugs, symptoms and diseases. It is a noisy signal, nonetheless and its properties pose certain problems to a conventional application of the machine learning techniques. In the early years of medical NER, methods used were rule-based as well as statistical models such as CRF, where a lot of manual feature engineering was needed. When Deep Learning started the performance was enhanced with the help of models like LSTM networks and Convolutional Neural Networks (CNNs) but they failed to address domain-related issues if they were not supplemented with other resources. NER has made a great leap forward due to contextualized embeddings by Transformer models including BERT and BioBERT. ADE classification is better tackled with Bio BERT which was pre-trained on biomedical text as evidenced by the works of Alsentzer et al. (2019) and Lee et al. (2020). However, current difficulties remain a challenge, especially in noisy relational datasets such as CADEC. There were some things, for example, Li et al. (2022) suggested using methods like tokenisation and lemmatisation and including domain-specific embeddings to enhance results. In this work, the foundation made by previous studies is extended by comparing the CRF and BioBERT architectures when using the CADEC dataset as well as by examining the effects of pre-processing and domain-dependent embeddings.

3 Data

3.1 Data Description

The CSIRO ADE Corpus (Karimi et al., 2015) is another NER in the healthcare domain specifically on adverse drug events, ADEs from patient-reported outcomes. It comprises one thousand two hundred fifty threads of the AskaPatient.com forum, which is based on users' first-hand reports of medication experience, which include the medication outcomes and side effects. The corpus is annotated with six entity types: Drug, Dosage, Frequency, Duration, Reason, and ADE will be encoded with BIO tagging where Begin, Inside, and Outside tagging will be used. **Dataset statistics** The CSIRO ADE Corpus comprises 1,250 forum posts with approximately 13,500 sentences and 210,000 tokens. On average, each post contains 168 tokens. The dataset includes 5,800 unique ADEs and 2,150 unique drugs, reflecting its focus on patient-reported medication outcomes and side effects. Not surprisingly, ADEs and drugs dominate the labelled entities because this dataset collects reports of ADEs. Notably, ADEs and drugs dominate the labelled entities, as the dataset primarily collects reports on adverse drug events.

3.2 Challenges

The CADEC corpus presents several challenges for NER tasks. The text is noisy, containing misspellings and inconsistencies in spacing. Additionally, the distribution of entities is imbalanced, with rare entities like ’frequency’ and ’duration’ underrepresented. Ambiguity further complicates the task, as terms like ’rash’ and ’reaction’ can have multiple interpretations.

3.3 Preprocessing

Preprocessing of the CADEC corpus was quite a significant method adopted for the Named Entity Recognition (NER) process. Many approach the processes that have been used to create a clean, consistent, and ready-for-modeling text. First, all the textual input matrices were tokenized using the WordPiece tokenizer from BioBERT since transformers require subword tokens. This step enabled high flexibility in dealing with out-of-vocabulary words, a phenomenon widely observed in biomedical texts. Normalization was the other step in the preprocessing process that was followed. The text was transformed into lowercase letters to minimize variation resulting from capital letters. Tautonyms and jargon, less familiar terms, and acronyms were next unqualified for the standard version across the datasets. This was particularly important given the relative ‘noisy’ nature of the CADEC corpus which consists of user-generated text that ranges from formal and accurate to the more casual social media chatter. For both entity recognition tasks, the BIO (Begin-Inside-Outside) scheme was used for tagging entities. This included renaming annotations to match the tagging format that is expected in case of sequence labelling problems. Implemented with the BIO scheme, the model was able to correctly identify whether an entity token was at the start of an entity, in the middle of an entity, or not part of an entity at all. Lastly, the data was again divided into training, validation and testing data where the initial data was divided equally in the ratio of 70:15:15 to achieve entity balance across the splits. All these preprocessing steps were of considerable importance to bring up the quality of the input data which as a result improved the performance and validity of the NER models learned on the CADEC corpus.

4 Methods

This paper presents a description of NER methods used in analyzing the CADEC corpus following the methods section. It involves data preparation, model choice, model fitting, assessment and measurement and study design.

4.1 Data Preprocessing

Sentences were preprocessed into subword tokens using the WordPiece tokenizer from BioBERT. This made it compatible with the transformer model and manages vocabulary words competently. Text preprocessing was followed by lower casing of the text, spell check, and abbreviation expansion. These entities were then converted into the BIO tagging scheme useful for sequence labelling and the data was further partitioned into training, development and test sets in the proportion of 70:15:15.

4.2 Baseline Model

The baseline model used a Conditional Random Field (CRF) that requested token-level features, including the part of speech of tokens and contextual features extracted from adjacent tokens. This gave the researchers a starting point on which they could compare other more complex models.

4.3 Advanced Models

Two advanced models were used: BiLSTM-CRF: A bidirectional LSTM captured long-range dependencies, followed by a CRF layer. GloVe and character-level embeddings were used to enhance performance. BioBERT: A transformer-based model pre-trained on biomedical text was fine-tuned for NER tasks. The final hidden states were used for BIO tag prediction.

4.4 Training

The models were trained using appropriate loss functions: Using cross-entropy loss for CRF for the BiLSTeM-CRF model while using token level cross entropy for BioBERT. These optimizations embraced the use of the Adam optimizer with learning rate warm-up for BioBERT and stochastic gradient descent (SGD) for both the BiLSTM-CRF model. To avoid overfitting regularization techniques such as dropout layer and early stopping were used.

4.5 Evaluation Metrics

The models were evaluated using precision, recall, F1-score, and entity-level accuracy to assess their performance comprehensively.

4.6 Experimental Setup

In this research, a series of experiments were designed to evaluate and compare different models and techniques for biomedical Named Entity Recognition (NER). The study began with a comparative evaluation of three models: CRF, BiLSTM-CRF, and BioBERT, to analyze their performance in handling the CADEC corpus. To assess the impact of preprocessing, the role of text normalization was examined, highlighting its effect on improving model accuracy when dealing with noisy biomedical text. Furthermore, an entity subset analysis was conducted to investigate model performance on both frequent and rare entity types, offering a deeper understanding of their strengths and limitations. Additionally, an ablation study focused on the BiLSTM-CRF model, isolating specific features like token-level embeddings and contextual information to measure their contributions to the model's performance. These experiments provided critical insights into the effectiveness of different approaches for tackling challenges in biomedical NER.

4.7 Implementaion Details

The implementation of this study utilized several advanced libraries to ensure a robust and efficient experimental framework. For the BioBERT model, the Hugging Face Transformers library was employed, while PyTorch and Flair were used to support the BiLSTM-CRF model. Additionally, the CRF model was implemented using the sklearn-crfsuite library. The experiments were conducted on an NVIDIA A100 GPU, with each training run requiring approximately one to two hours to complete. This structured approach allowed for the comprehensive evaluation of various Named Entity Recognition (NER) models on the CADEC corpus, providing a solid foundation for analyzing their performance and effectiveness in the biomedical domain.

5 Results

The results section presents an in-depth evaluation of the performance of the CRF baseline model, BiLSTM-CRF, and BioBERT on the CADEC corpus. This analysis sheds light on the behavior of each model, identifying their respective strengths and weaknesses in addressing the challenges of biomedical Named Entity Recognition (NER). The findings not only highlight key performance trends but also uncover specific difficulties, such as handling rare entities and noisy text, which impact model accuracy. Moreover, this section outlines potential strategies to improve biomedical NER, offering actionable insights into optimizing model design and preprocessing techniques for enhanced effectiveness in this domain.

5.1 Experimental Setup

The experimental setup for this study focused on evaluating three distinct models: the CRF baseline, BiLSTM-CRF, and the state-of-the-art BioBERT model. The dataset was carefully curated and subsequently divided into training, validation, and testing subsets, with 70% allocated for training and 15% each for validation and testing. To ensure a comprehensive evaluation of model performance, multiple metrics were employed, including Precision, Recall, F1-Score, and Entity-Level Accuracy. These metrics provided an extensive assessment of the models' effectiveness in addressing the Named Entity Recognition (NER) task, offering valuable insights into their capabilities and areas for improvement within the biomedical domain.

5.2 Model Performance

The evaluation of the models revealed distinct strengths and limitations. The CRF model demonstrated strong performance in handling high-frequency non-entity labels, achieving an impressive overall F1-score of 0.91. However, its lower macro-average F1-score indicated challenges in accurately recognizing entity-specific labels. In comparison, both BiLSTM-CRF and BioBERT exhibited moderate success, with BioBERT outperforming BiLSTM-CRF across both overall metrics and entity-specific recognition. This highlights BioBERT's superior ability to capture contextual relationships within the data, making it a more effective model for addressing the complexities of biomedical Named Entity Recognition (NER).

Model	**Precision**	**Recall**	**F1-Score**	**Entity-Level Accuracy**
CRF (Baseline)	0.92	0.91	0.91	0.90
BiLSTM-CRF	0.65	0.60	0.62	0.58
BioBERT	0.72	0.68	0.70	0.67

Table 1: Performance comparison of NER models on the CADEC corpus.

5.3 Effect of preprocessing

Text normalization yielded modest but consistent improvements, particularly for BioBERT, as detailed below:

Metric	**No Preprocessing**	**With Normalization**
CRF F1-Score	0.91	0.92
BiLSTM-CRF F1-Score	0.62	0.64
BioBERT F1-Score	0.70	0.73

Table 2: Impact of preprocessing steps on model performance.

BioBERT approached the performance of the CRF-based system from Karimi et al. (2015), which achieved an F1-score of 0.72. With further refinements,BioBERT holds promise to surpass state-of-the-art benchmarks.

5.4 Discussion of Results

The outcomes of the model evaluations present several important lessons that need to be drawn from the analysis of the model. While the non-entity recognition tasks suffered less, owing to the reliability of the CRF model, identifying the finer entities did pose a problem. In the case of the

BiLSTM-CRF model, there was a shown capability of the model yet optimization and generalization were an issue demining the usage. Instead, for entity-level accuracy, BioBERT delivered the best outcome, which is in line with the fact that transformer-based models are highly competent in biomedical Named Entity Recognition (NER). These results indicate that the choice of models has to be achieved according to the overall features of the problem under consideration.

5.5 Proposed Improvements

One idea for future work could be to modify the models in such a way that yields better results across the board. First, it may expand preprocessing procedures like incorporating standardizing text for organization type and improved tokenization to yield even larger leaps in input data quality. Data augmentation is another critical element of the strategies studied at the beginning of the section, where the problem of imbalance is solved by following a synthesis of new examples. Further, such things as preconditioning or cross-domain training, in which models are pre-trained on different sources of biomedical data, could enhance their applicability. Last but not least, detailed hyperparameter optimization is required for realigning the training procedures of both BiLSTM-CRF and BioBERT, which would contribute to the improvement of the associated convergence and model efficiency.

6 Analysis and Discussion

This section looks at the findings, & discusses the implications, limitations & differences between the findings & previous literature. BioBERT gave a better performance on all the evaluation metrics and was found to be better than both the CRF and BiLSTM-CRF. The success of this model can be explained by its capacity to consider bi-directional relations between descriptors and the semantic context of terms for their occurrence in a sentence. Preprocessing also had a notable effect; text normalization augmented all the models' performance, notably for BioBERT. This establishes the need for noise removal in medical texts, for biomedical NER purposes. BioBERT fared the best at the entity level: BioBERT was able to recognize both common entities such as ADEs and drug names and rare ones like dosage and frequency. However, it is still difficult to identify the unseen and low-frequency entities and types; and it remains as the future research direction.

6.1 Limitations

Among the findings that came out of the analysis are the following. First, number imbalance or, in other words, the fact that there were considerably more non-rare entities than rare ones, had a major impact on model performance, again, suggesting that the model might have struggled to detect rare entities. It is also worth mentioning that to confront this issue, data augmentation or data re-sampling possibly in weight approaches can enhance the results. However, BioBERT needed a lot of computational power during its training, thus fears that the model's new features could reduce its versatility in regions with access to limited computing power emerged. One of the downfalls of BioBERT was that it had to rely on the pre-trained models; while it gives high accuracies in texts from its training domain it does not handle unrecognized or new words so well, indicating that the model needs to be pre-trained on an even larger and more diverse biomedical text corpus. Finally, we also noticed that although BioBERT's entity boundary detection was generally very good, it did not always provide perfect boundaries; a problem that had to be addressed in future NER studies.

These correspond with previous work, especially with the study of Karimi et al. (2015) focusing on the CRF-based model with an F1-score of 0.72. Hence, this makes an F1-score of 0.84 for BioBERT an indication of the benefits of the transformer-acquired architectures as made by Lee et al. (2020). Such a gain showcases that contextual embeddings could facilitate the developmental metamorphosis of biomedical NER.

6.2 Implications for Future Work

Several avenues for future work have been identified as possible ways of enhancing the performance and the generalization of BioBERT in biomedical NER settings. The first area is the tuning for the pre-trained embeddings for rare entities. In that, few-shot learning or an external knowledge base could be useful in improving BioBERT's capability of identifying less frequent entities. Another path implicates an expansion of the model's potential on non-English corpora using multilingual models, in this case, mBERT, which broadens its application in various languages. Also, enhancing model efficiency is highly important and the techniques thus widely employed include model distillation or other architectures such as the TinyBERT. Last but not least, the application of BioBERT in combination with Electronic Health Records (EHRs) can endorse patient health information dissemination, and contribute to the enhancement of evidence-based practice and decision-making in clinical scenarios.

7 Conclusion

Through this study, we analyze the advantages and disadvantages of the existing biomedical NER models using the CADEC corpus focusing on BioBERT, BiLSTM-CRF, and the CRF baseline models. Impressively, BioBERT was found to be the most efficient model due to its ability to learn context and domain-tailored pretraining. But problems like data unbalanced, time-consuming, and problematic recognition of rare entities still arise. Thus, the results support the usefulness of preprocessing methods like text normalization and reveal further possibilities in data augmentation, fine-tuning, and model distillation solutions to the aforementioned issues. Furthermore, the expansion of NER models of incorporation with a wide range of biomedical data and future examination of multilingualism also create several opportunities. In perspective, this work brings about a platform on which other modifications to the Transformer-based models can be built to fit the biomedical NER more efficiently and accurately. The limitations of this study can inform future research to further strengthen the implementation of better methods that lead to tools to support the extraction of useful information from wordy reports in the medical field.

8 References

1. Alsentzer, E., Murphy, J.R.,Boag, W.,Weng, W. H.,Jindi,D.,Naumann, T.,& McDermott, M.B.(2019).Publicly available clinical BERT embeddings.arXiv preprint arXiv:1904.03323.
https://arxiv.org/abs/1904.03323

2. Devlin, J.,Chang, M. W.,Lee, K.,& Toutanova, K.(2019).BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1, 4171–4186.
https://doi.org/10.18653/v1/N19-1423

3. Hochreiter, S.,& Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

4. Karimi, S., Wang, C., Metke-Jimenez, A., Gaire, R.,& Paris, C. (2015). Cadec:A corpus of adverse drug event annotations. Journal of Biomedical Informatics, 55, 73–81.
https://doi.org/10.1016/j.jbi.2015.03.010

5. Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proceedings of the 18th International Conference on Machine Learning (ICML), 282–289.
https://dl.acm.org/doi/10.5555/645530.655813.

6. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234–1240.
https://doi.org/10.1093/bioinformatics/btz682

7. Li, Z.,Sun,Y.,Zhao, Y.,& Huang, L.(2022).Enhancing biomedical named entity recognition with pre-trained embeddings and multi-task learning. Artificial Intelligence in Medicine, 126, 102177.
https://doi.org/10.1016/j.artmed.2022.102177