Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220615

Named Entity Recognition in Pubmed Abstracts for Pharmacovigilance Using Deep Learning

T. Trang NGHIEM[a] and Cedric BOUSQUET[b,c,1]
[a]Institute of Thermal, Mechanical and Material Sciences (ITheMM EA 7548), University of Reims Champagne-Ardenne, 51687 Reims, France.
[b]Unit of public health and medical informatics, CHU de Saint Etienne, France
[c]Sorbonne Université, Inserm, université Paris 13, Laboratoire d'informatique médicale et d'ingénierie des connaissances en e-santé, LIMICS, F-75006 Paris, France

Abstract. Methods of natural language processing associated with machine learning or deep learning can support detection of adverse drug reactions in abstracts of case reports available on Pubmed. In 2012, Gurulingappa et al. proposed a training set for the recognition of named entities corresponding to drugs and adverse reactions on 3000 Pubmed abstracts. We implemented a classifier using deep learning with a Bi-LSTM and a CRF layer that achieves an F-measure of 87.8%. Perspectives consist in using BERT for improving the classifier, and applying it to a large number of Pubmed abstract to build a database of case reports available in the literature.

Keywords. Artificial intelligence, pharmacovigilance, Pubmed, Deep learning

1. Introduction

In order to monitor drug safety with the medical literature, it is necessary to carry out bibliographic queries on a regular basis. However, pharmacovigilance teams have only Pubmed and its user interface which is intended for all users. Our aim is to build a database consisting of Pubmed abstracts related to adverse drug reactions (ADR). We employ the training set based on 3000 Pubmed abstracts on case reports likely to describe ADRs implemented by Gurulingappa et al. [1]. Most previous works have applied machine-learning-based methods, but not deep learning [2, 3].

Our first step as described in this paper was to implement a classifier to detect named entities related to drugs and ADRs using a conditional random field (CRF) and a Long short-term memory (LSTM) deep learning layers. Our model can be trained quickly (less than 1 hours) on only one 16 GB RAM computer. We compared our results with those of other recent works using named entity recognition of drugs and ADRs with deep learning [4, 5, 6].

[1] Corresponding Author, Dr Cedric Bousquet, SSPIM, Bâtiment CIM42, chemin de la Marandière, Hôpital Nord, 42055 Saint Etienne; E-mail: cedric.bousquet@chu-st-etienne.fr.

2. Methods

We used a set of 3000 Pubmed abstracts annotated in 2012 by Gurulingappa *et al.* that contained 6821 sentences with a relation between a drug and an adverse reaction, and 16,695 sentences without [1]. We implemented a CRF with sklearn-crfsuite and two versions of a Bi-LSTM deep learning model with or without a CRF layer. Embeddings of size 40 were fed into the network using Keras' embedding layer. We used an IO tagging of tokens in Pubmed abstracts where "I" corresponds to "Inside", including two types of "I" (I_ADR and I_DRUG), and "O" stands for "Out". Each experiment was repeated five rounds, and the averaged results were taken for all metrics (precision, recall and F1-score). Obtained results are reported below with those of previous publications.

3. Results and Discussion

Table 1 shows the results we obtained with our models (first three lines), and results obtained by other authors (next three lines).

Table 1. Precision, recall and F1-Score for our models and other publications. POS stands for Part-of-speech in Named Entity Recognition model.

Models	Precision (%)	Recall (%)	F1-Score
CRF (with POS)	92.8	81.1	86.6
Bi-LSTM	85.7	85.6	85.6
Bi-LSTM + CRF	92.6	83.5	87.8
Ramamoorthy et al. [4]	88.4	82.4	85.3
P. Ding et al. [5]	86.7	94.8	90.6
Hussain et al. [6]	98.2	96.4	97.6

Our lightweight model is capable to extract quite efficiently named entities on drugs and adverse reactions from Pubmed abstracts. We plan to detect ADRs on a large number of case reports to build a reference database to improve queries for pharmacovigilance in the literature. A Bidirectional Encoder Representations from Transformers (BERT) model [7] should be implemented to improve the performances of our classifier.

References

[1] Gurulingappa, H, Mateen-Rajpu A, Toldo L. Extraction of potential adverse drug events from medical case reports. J biomedical semantics. 2012;3(1):1-10.

[2] Sarker A, Ginn R, Nikfarjam A, O'Connor K, Smith K, Jayaraman S, Upadhaya T, Gonzalez G. Utilizing social media data for pharmacovigilance: A review. J Biomed Inform. 2015;54:202-12.

[3] Luo Y, Thompson WK, Herr TM, Zeng Z, Berendsen MA, Jonnalagadda SR, Carson MB, Starren J. Natural Language Processing for EHR-Based Pharmacovigilance: A Structured Review. Drug Saf. 2017;40(11):1075-1089.

[4] Ramamoorthy S, Murugan S. An attentive sequence model for adverse drug event extraction from biomedical text. arXiv preprint. 2018. arXiv:1801.00625.

[5] Ding P, Zhou X, Zhang X, Wang J, Lei, Z. An attentive neural sequence labeling model for adverse drug reactions mentions extraction. IEEE Access. 2018; 6:73305-15. doi: 10.1109/ACCESS.2018.2882443

[6] Hussain S, Afzal H, Saeed R, Iltaf N, Umair MY. Pharmacovigilance with Transformers: A Framework to Detect Adverse Drug Reactions Using BERT Fine-Tuned with FARM. Comput Math Methods Med. 2021;2021:5589829. doi: 10.1155/2021/5589829. eCollection 2021.

[7] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint. 2018. arXiv:1810.04805.