Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2019) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

Postproceedings of the Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2020

On the accuracy of different neural language model approaches to ADE extraction in natural language corpora

Alexander Sboev[a,b,1], Anton Selivanov[a], Gleb Rylkov[a], Roman Rybka[a]

[a]*National Research Centre "Kurchatov Institute", Moscow, Russia*
[b]*National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), Moscow, Russia*

Abstract

The problem of extracting mentions of adverse events and reactions from text is especially relevant nowadays due to rapid emergence of datasets including such events, and progress in text analysis tools. This paper presents a comparison of existing methods for the task of automated extraction of adverse events from natural language texts. The considered methods are based on neural-network language models, pre-trained on different sets of unlabeled data. Experiments have been performed on the n2c2-2018 and CADEC corpora, using metrics coined within the CoNLL competition. Models of the aforementioned type show efficient solution of this task, provided sufficient amount of labeled training samples during.

Keywords: natural language processing; artificial neural networks; language models; named entity recognition; adverse drug events; pharmacovigilance

1. Introduction

Accuracy of the up-to-date neuronet models for Adverse Drug Events (ADE) extraction from colloquial written speech, is a question in the spotlight of machine learning due to its importance for the tasks of pharmacovigilance and post-clinical drug investigation. As mentioned at National NLP Clinical Challenge 2018 (n2c2-2018), automatic extraction of the entities of such type is a challenging task of high importance and a way for further improvement of natural language processing (NLP) of medical texts [8]. There currently exist a number of corpora which may be considered canonical for evaluating this accuracy [14, 7], such as n2c2, MADE, PsyTAR, TwiMed, CADEC. At the same time, new neural-network-based models have been recently put into practice, called language models [5], that are based on advanced deep learning topologies containing transformer [18] layers and trained preliminarily on huge

[1] Corresponding author. e-mail: Sboev_AG@nrcki.ru

unlabeled corpora. However, the accuracy such models can achieve depends on a multitude of factors: the size of the training corpus, complexity of its language and markup, and its richness with entities related to adverse effects, in particular, the ratio of texts containing such entities to texts that don't contain. The current paper strives to assess the effect of these factors on the ADE recognition accuracy.

With this purpose, the paper aims at a comparative study of the existing high-performing algorithms in order to select those that would be most efficient for further analysis of other corpora with their structure analogous to the canonical ones above, but comprising texts in other languages, in particular, Russian texts. In addition, we examine differences in accuracy between a corpus of colloquial texts from the Internet and one of medical records written by specialists. The models used in our comparison are variations of pre-trained deep language models that differ by a few quantitative parameters and by data they were pre-trained on; these are described briefly in Section Methods. The corpora on which we have performed training and accuracy evaluation are described in Section Corpora. In Section Experiments, accuracies of the models are presented and compared.

In this study we compare language models pre-trained on different datasets. The models used are the architectures of Bidirectional Encoder Representations from Transformers (BERT) [5], which involve attention mechanisms (multi-head attention mechanism) within their Transformer layers [18]. Such architectures proved to be promising for various natural language understanding (NLU) tasks, including named entity recognition (NER). Pre-training the models is performed on unlabeled data, analogously to supervised learning for the tasks of 1) reconstructing hidden tokens, 2) determining whether two sentences come successive in a text. In the literature, there are studies of these architectures [7] while using different data (different both by amount and by composition), number of layers, and hyperparameters. Thus, our research is to select the most efficient model out of those existing to date.

BlueBERT (M+P) [14] is a model based on the BERT language model, trained on the BooksCorpus (800M words) [20] and the English Wikipedia (2,500M words), and later trained on a combination of medical texts: PubMed abstracts (approx. 4000M words) and de-identified clinical notes MIMIC-III (approx. 500M words) [10]. That training consisted of 5M steps on the PubMed corpus and 0.2M steps on the MIMIC-III corpus. We use the basic variant of the model (8 Transformer layers, maximum token sequence length of 512).

BioBERT [13] is also based on BERT trained on BooksCorpus and English Wikipedia, but subsequently trained on paper abstracts from PubMed (4,500M words) during 1M steps.

EnDrBERT [17] is a model based on Multilingual BERT, pretrained on a corpora of Wikipedia texts in 104 languages, and then trained on a corpora of user comments from popular forums such as webmd.com, askapatient.com, drugs.com, dailystrength.org, patient.info, amounting to the total of 2.6M texts and 254M tokens (separate words, numbers, or punctuation marks). The latter training was performed with hyperparameters the same as the original ones of Multilingual BERT.

PubMedBERT [7] differs from the previously described models: in PubMedBERT, the BERT model was trained solely on texts of biomedical domain from PubMed, without the use of other corpora like WordCorpus and Wikipedia. We use two variants of this model, called "PubMedBERT-abstracts" and "PubMedBERT-full-text". The first model was trained on paper abstracts (at least 128 word long, approx. 14 million abstracts, 3.2 billion words, 21 GB). Model was trained for 62,500 steps with a batch size of 8,192. The second variant of the model was trained on full texts of papers from PubMed Central (PMC), with the total volume of pretraining data increased substantially to 16.8 billion words (107 GB). The pretraining process was extended to 100K steps in total.

XLM-Roberta-large [2] has the architecture of BERT Large [5] in its foundation. Training was performed (without any other pre-training) on 2 TB of text data in 100 languages (including rare languages) from the CommonCrawl project. This model showed an improvement over the original Multilingual BERT.

Table 1 is a brief summary of the complexity of the models used, presented in the common format [2], it also describes the data used for pre-training of the models (PMa denotes PubMed abstracts, PMC — PubMed Central papers). There, #lgs is the number of languages in the data on which the models were pre-trained, L is the number of layers, H_m is the number of hidden states of the model, H_ff is the dimension of the feed-forward layer, A is the number of attention heads, V is the size of the vocabulary, and #params is the total number of trainable model parameters. For Transformer encoders, the number of parameters can be approximated by $4LH_\mathrm{m}^2 + 2LH_\mathrm{m}H_\mathrm{ff} + VH_\mathrm{m}$.

The neural-network-based language models listed above are considered promising for a study in regard to the current task. For more detailed description of these models, as well as the approaches for tokenizing and building the vocabulary, we refer the reader to the respective original papers.

Table 1. The number of texts and ADE entities in the training, validation (denoted as Dev), and testing sets of the two corpora used

Model	#lgs	data	L	H_{m}	H_{ff}	A	V	#params
BlueBERT-base	1	BooksCorpus, Wikipedia + PMa, MIMIC-III	12	768	3072	12	30k	110M
BioBERT-base	1	BooksCorpus, Wikipedia + PMa	12	768	3072	12	30k	110M
EnDR-BERT	104	Wikipedia + Medical web-resources	12	768	3072	12	110k	172M
PubMedBERT abstracts	1	PMa	12	768	3072	12	30k	110M
PubMedBERT fulltext	1	PMC	12	768	3072	12	30k	110M
XLM-RoBERTa	100	CommonCrawl	24	1024	4096	16	250k	550M

2. Corpora

2.1. CADEC [12]

CSIRO Adverse Drug Event Corpus (CADEC) is an annotated corpus of medical forum posts on patient-reported Adverse Drug Events (ADEs). The corpus is sourced from posts on social media, and contains text that is largely written in colloquial language and often deviates from formal English grammar and punctuation rules. The quality of the annotations is ensured by annotation guidelines, multi-stage annotations, measuring inter-annotator agreement, and final review of the annotations by a clinical terminologist. Corpora was split in the proportion of 72%/8%/20% (by the number of words) into training, validation, and testing sets, the same way as in [15].

2.2. n2c2-2018 [8]

n2c2 2018 is a dataset from the National NLP Clinical Challenge of the Department of Biomedical Informatics (DBMI) at Harvard Medical School. The dataset contains clinical narratives, and builds on past medication extraction tasks, but examines a broader set of patients, diseases, and relations as compared with earlier challenges. One of the subtasks of the challenge was: “Can NLP systems automatically discover adverse event in clinical narratives?” Train and test sets determined according to the challenge with further split of train set into training and validation sets on a 90/10 ratio.

The number of texts and ADE entities in these sets is in Table 2.

Table 2. The number of texts and ADE entities in the training, validation (denoted as Dev), and testing sets of the two corpora used

Feature	CADEC			n2c2-2018		
	Train set	Dev set	Test set	Train set	Dev set	Test set
Number of texts	845	92	311	274	29	202
Number of texts with ADE entities	749	78	277	215	21	154
Number of texts w/o ADE entities	96	14	34	59	8	48
Number of words	82139	9110	30407	635806	70737	463509
Number of ADE entities	4051	459	1427	889	70	625
Avg text length	97.21	99.02	97.77	2320.46	2439.21	2294.6
Min text length	2	3	2	102	183	112
Max text length	480	840	615	5455	7838	7202

3. Experiments

Models' performance is assessed by the entity recognition quality metric from CoNLL-2003 Shared Task [16], the F1-score:

$$F1 = \frac{2 * P * R}{P + R},$$

where precision P is a part (in percentage) of correct predictions of the entities in model outputs, and recall R is a part of the entities recognized correctly by the model. An entity is counted as recognised correctly if its beginning and ending positions in the text match exactly what the model outputs.

We used BIO labeling scheme for named entity recognition tasks, each word of a text has one label, and language models trained and used the same way as for sequence processing, with tag prediction for each element of the sequence.

Table 3 presents F1-scores for experiments with fine-tuning of pre-trained language models on CADEC and n2c2 datasets. There were two experiment sets: models from the first one trained to predict all tag types from initial dataset ("Trained on all entities type"), models from the second one trained as binary classifiers for ADE tag, other entities were treated as words with no tag ("Trained only on ADE entities").

Table 3. ADE-entities extraction accuracy (F1-conll, %)

No.	Neural network model	Trained only on ADE entities		Trained on all entity types	
		n2c2	CADEC	n2c2	CADEC
1	BlueBERT base (M+P)	44.4	67.51	47.44	68.43
2	BioBERT base	40.72	66.6	48.13	67.1
3	EnDrBERT	35.28	67.94	46.77	68.52
4	PubMedBERT-abstracts	33.79	65.47	40.44	64.79
5	PubMedBERT-fulltext	42.78	65.34	31.51	65.29
6	XLM-Roberta-large	33.43	69.68	48.66	69.66

As it's shown in the table, fine-tuning on the whole tag set makes complex models to achieve higher accuracy of ADE entities extraction in comparison with fine-tuning on the ADE tags only. Further in this section experiments presented on influence of train set size on accuracy with both types of fine-tuning (with all tags and with ADE tags only).

The best results achieved with XLM-Roberta-Large model, f1-score of 48.66% and 69.66% for n2c2 2018 and CADEC, respectively. This model is the most complex from considered ones (see Table 3), it needs more examples for training and fine-tuning. It reasons low accuracy on n2c2 when only ADE tags are used. At the same time, with enough data, hyperparameter setting, and model fitting on domain-specific data, comparable results could be achieved with less complex models (see exp. 1 in Table 3).

Achieved results are comparable with worldwide experience: the best results on CADEC (with f1-conll metrics) are in range from 63% [6] to 70% [4]. n2c2 2018 Challenge used f1-partial (lentient) metric, which is 23% for baseline approach based on dictionary and word features, and conditional random field (CRF) model [19]; from 40% to 50% in average with use of various word vector representation and machine learning models (crafted features, ELMo, CRF, character CNN, BiLSTM, etc.) [1, 3, 9, 11, 19]. XLM-Roberta-Large achieves 48% on f1-partial metric on this dataset. State-of-the-Art is 56%, based on a BiLSTM-CRF model with weighted voting algorithm, feature set included ELMo pre-trained on MIMIC-III dataset, characters embedding based on a convolutional neural network, manually labeled information about text section, PoS-tags, and word shape information [8]. Therefore obtained results are consistent with the ones from the other researchers.

Figure 1 presents training set size influence on accuracy of ADE entities extraction on CADEC dataset. BlueBERT base (M+P) model was used as an example. There, selecting portions of the training set is performed in a way that

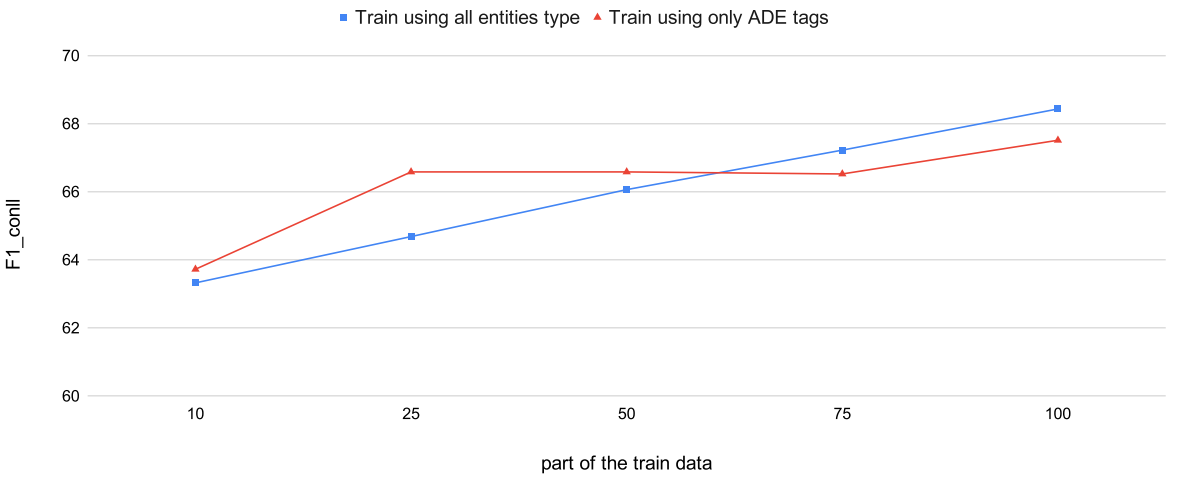

Fig. 1. Dependence of the accuracy of BlueBERT base (M+P) model on the CADEC dataset training set size.

each following part included all texts from the previous one. Validation and test sets were fixed. Results show that training set size increasing makes beneficial usage of the whole set of tags, demonstrating a more stable curve.

Validation and test sets were fixed. Results show that training set size increasing makes beneficial usage of the whole set of tags (hypothetically allows language model to extract internal relations between different entities or let it change weights in more sophisticated way to predict more tags at once).

Conclusion

Results achieved on two datasets of texts: clinical (n2c2-2018), and colloquial (CADEC), XLM-Roberta-Large has best accuracy for the models based on Transformer layers. It is shown that high complexity of such models makes it preferable to train them on the whole tag set of the corpus to achieve better accuracy for ADE extraction. This phenomenon is confirmed on both text datasets. The best achieved f1-score for ADE extraction is 69.66% for CADEC, repeating the best result of other researchers, and 48.66% for n2c2-2018. It may be explained by high number of texts with ADE entities in CADEC dataset (the part of texts with ADE is 88%, and the number of texts is relatively big). As for n2c2-2018, texts number is relatively small, and has lower part of ADE mentions, which leads to higher results variability. Further research aimed on model design for named entities extraction from pharmacological text on Russian.

Acknowledgements

This work has been supported by the Russian Science Foundation grant 20-11-20246 and carried out using computing resources of the federal collective usage center Complex for Simulation and Data Processing for Mega-science Facilities at NRC “Kurchatov Institute”, http://ckp.nrcki.ru/.

References

[1] Christopoulou, F., Tran, T.T., Sahu, S.K., Miwa, M., Ananiadou, S., 2020. Adverse drug events and medication relation extraction in electronic health records with ensemble deep learning methods. Journal of the American Medical Informatics Association 27, 39–46.

[2] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V., 2019. Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116 .

[3] Dai, H.J., Su, C.H., Wu, C.S., 2020. Adverse drug event and medication extraction in electronic health records via a cascading architecture with different sequence labeling models and word embeddings. Journal of the American Medical Informatics Association 27, 47–55.

[4] Dai, X., Karimi, S., Paris, C., 2017. Medication and adverse event extraction from noisy text, in: Proceedings of the Australasian Language Technology Association Workshop 2017, pp. 79–87.
[5] Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 .
[6] Ge, S., Wu, F., Wu, C., Qi, T., Huang, Y., Xie, X., 2020. Fedner: Privacy-preserving medical named entity recognition with federated learning .
[7] Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H., 2020. Domain-specific language model pretraining for biomedical natural language processing , arXiv:2007.15779.
[8] Henry, S., Buchan, K., Filannino, M., Stubbs, A., Uzuner, O., 2019. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. Journal of the American Medical Informatics Association 27, 3–12. URL: https://academic.oup.com/jamia/article-pdf/27/1/3/34152182/ocz166.pdf, doi:10.1093/jamia/ocz166.
[9] Huang, Z., Xu, W., Yu, K., 2015. Bidirectional lstm-crf models for sequence tagging. arXiv preprint arXiv:1508.01991 .
[10] Johnson, A.E., Pollard, T.J., Shen, L., Li-wei, H.L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., Mark, R.G., 2016. Mimic-iii, a freely accessible critical care database. Scientific data 3, 160035.
[11] Ju, M., Nguyen, N.T., Miwa, M., Ananiadou, S., 2020. An ensemble of neural models for nested adverse drug events and medication extraction with subwords. Journal of the American Medical Informatics Association 27, 22–30.
[12] Karimi, S., Metke-Jimenez, A., Kemp, M., Wang, C., 2015. Cadec: A corpus of adverse drug event annotations. Journal of biomedical informatics 55, 73–81.
[13] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J., 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36, 1234–1240. URL: https://academic.oup.com/bioinformatics/article-pdf/36/4/1234/32527770/btz682.pdf, doi:10.1093/bioinformatics/btz682.
[14] Peng, Y., Yan, S., Lu, Z., 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets, in: Proceedings of the 18th BioNLP Workshop and Shared Task, pp. 58–65.
[15] Stanovsky, G., Gruhl, D., Mendes, P., 2017. Recognizing mentions of adverse drug reaction in social media using knowledge-infused recurrent models, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pp. 142–151.
[16] Tjong Kim Sang, E.F., De Meulder, F., 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition, in: Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4, Association for Computational Linguistics. pp. 142–147.
[17] Tutubalina, E.V., Miftahutdinov, Z.S., Nugmanov, R.I., Madzhidov, T.I., Nikolenko, S.I., Alimova, I.S., Tropsha, A.E., 2017. Using semantic analysis of texts for the identification of drugs with similar therapeutic effects. Russian Chemical Bulletin 66, 2180–2189. URL: http://link.springer.com/article/10.1007/s11172-017-2000-8, doi:10.1007/s11172-017-2000-8.
[18] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I., 2017. Attention is all you need, in: Advances in Neural Information Processing Systems, Curran Associates, Inc.. pp. 5998–6008. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
[19] Wei, Q., Ji, Z., Li, Z., Du, J., Wang, J., Xu, J., Xiang, Y., Tiryaki, F., Wu, S., Zhang, Y., et al., 2020. A study of deep learning approaches for medication and adverse drug event extraction from clinical text. Journal of the American Medical Informatics Association 27, 13–21.
[20] Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., Fidler, S., 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books, in: The IEEE International Conference on Computer Vision (ICCV).