MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI231049

Influence of Context in Transformer-Based Medication Relation Extraction

Luise MODERSOHN[a,b,1] and Udo HAHN[a]
[a]*Jena University Language & Information Engineering (JULIE) Lab, Friedrich-Schiller-Universität Jena, Jena, Germany*
[b]*AIIM, TU München, München, Germany*
ORCiD ID: Luise Modersohn https://orcid.org/0000-0002-8258-4366, Udo Hahn https://orcid.org/0000-0002-5052-0245

Abstract. The extraction of medication information from unstructured clinical documents has been a major application of clinical NLP in the past decade as evidenced by the conduct of two shared tasks under the I2B2 and N2C2 umbrella. We here propose a new methodological approach which has already shown a tremendous potential for increasing system performance for general NLP tasks, but has so far not been applied to medication extraction from EHR data, namely deep learning based on transformer models. We ran experiments on established clinical data sets for English (exploiting I2B2 and N2C2 corpora) and German (based on the 3000PA corpus, a German reference data set). Our results reveal that transformer models are on a par with current state-of-the-art results for English, but yield new ones for German data. We further address the influence of context on the overall performance of transformer-based medication relation extraction.

Keywords. Clinical natural language processing, transformers, relation extraction

1. Introduction

Due to its tremendous clinical relevance, the extraction of medication information from unstructured clinical documents has been a major application of clinical natural language processing (NLP) in the past decades. This is evidenced by two shared tasks focusing on medication extraction (with over-lapping though slightly differing informational attributes) within the framework of I2B2 [1] and N2C2 [2]. The most successful teams in these shared task competitions used, at that time, advanced methodological approaches – a mixture of rule-based and statistical classifiers based on feature engineering for I2B2 and BiLSTM-CRF neural networks for N2C2.

Since the latest N2C2 challenge on medication extraction, methodological advances in the field of general NLP (mostly working on newspaper and Wikipedia articles) are based on transformer models [3], which typically outperform non-transformer models by a large margin due to their built-in attention mechanism. In this paper, we investigate whether transformers also perform well for medication extraction from EHR data (for a survey and comparison of current clinical transformer architectures, cf. [4]).

[1]Corresponding Author: Luise Modersohn, AIIM, Klinikum rechts der Isar der TU München, Ismaninger Str. 22, 81675 München, Germany.

In the following sections, we will give an overview of previous work on automatic medication-related information extraction for both drug-related named entities and relations. We then describe our data sets and experiments, and, finally, present and discuss our evaluation results.

2. Methods

In our approach, there is only one method to automatically extract both medication-related named entities and relations. It is based on the transformer model[2] [4,3] and its implementation in BERT [5]. Training a BERT-based language model from scratch is a time-consuming and resource-intensive task that requires a lot of training data. Fortunately, trained language models are already available for many different languages. Thus, we only need to fine-tune pre-trained language models on our clinical corpora.

2.1. Datasets

2.1.1. I2B2 2009

The Third Informatics for Integrating Biology and the Bedside (I2B2) challenge 2009 [1] was the first dealing exclusively with automatic medication information extraction. It provided clinical discharge summaries with annotated gold data, as well as detailed annotation guidelines. The best system scored with an overall macro F1-score of 0.857. For our work, we used both the annotated data originally provided by the I2B2 team, as well as the community-annotated dataset [6]. The entire collection consists of 260 documents. For more details about the entities and relations of the I2B2 2009 dataset and their distribution, see Figure 1.

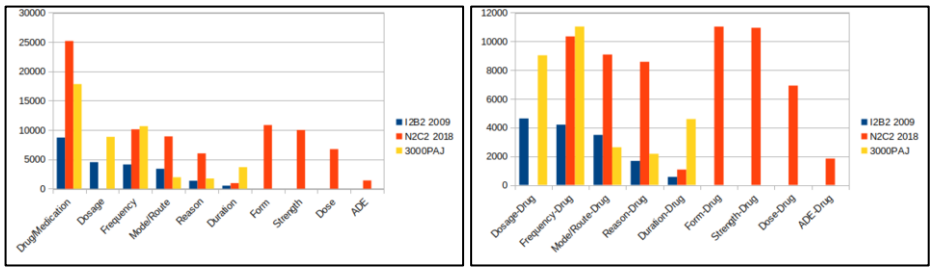

Figure 1. Distribution of the relevant entity (left) and relation (right) types in all three annotated corpora.

This dataset does not provide explicit relations, because its annotations consist of a medication and several relevant attributes associated with the specific administered drug, such as frequency, mode or dose, and their respective offsets. Thus, relations are given implicitly only. We named them according to their entity pair (frequency-drug, dose-drug, etc.), as there are no predefined labels available in the original dataset.

[2] For a more high level introduction, see Tunstall, L., von Werra, L., Wolf, T.: Natural Language Processing with Transformers: Building Language Applications with Hugging Face. O'Reilly Media, Inc.. 2022

2.1.2. N2C2 2018

The First track of the National NLP Clinical Challenges (N2C2) competition in 2018 [2] picked up but also modified many ideas from the I2B2 2009 medication challenge. The initial medication(-related) entities and relations were refined and an important complementary aspect of drug use, adverse drug events (ADE), was added in this challenge. The best system in the entity categories achieved a strict macro average F1-score of 0.89. For relation classification a macro average F1-score of 0.96 was reported. The dataset consists of 505 discharge summaries with both entity and relation annotations. Figure 1 provides more information about the entity and relation types, as well as their respective counts and ratios.

2.1.3. $3000PA_J$

The 3000PA data set is currently one of the largest German clinical text corpora [7]. However, as 3000PA is a corpus physically distributed at three university hospital sites, we will only use the local part from the Jena University Hospital of this corpus, termed $3000PA_J$, to which we have direct access (no allowances for external use are in place). This part contains 1,106 German discharge summaries. For a detailed description of the entity and relation types and their respective number of occurrences, see figure 1.

The guidelines we used to annotate this German corpus are based on the I2B2 2009 annotation guidelines. Thus, there is an intrinsically high consensus on the definition of entity and relation types.

2.2. Experiments

We will now describe our experimental setup in more detail. Our initial experimental code base was a clone of the PyTorch version of BioBERT.[3] In our experiments, we used the transformers library huggingface[4] for model training. The hyperparameter optimization was performed using the BOHB implementation of the Microsoft/NNI framework.[5] For each dataset, we performed the same number of steps.

To generate the train-dev-test splits, we generally used a 65:10:25 distribution. If a predefined split was already set, e.g., as for the N2C2 2018 dataset, we kept the test split as is and only used 10% of the train split as dev. Notably, we did not exploit the train-test split of the I2B2 2009 dataset, since that train split consists of only 10 documents, compared to the 250 community-annotated documents in the test split. Thus, we decided to group them together and split them according to the 65:10:25 distribution from above.

For the English-speaking language community, there are many transformer-based SOTA language models available – not only generic, but also domain-specific ones. In our case, we needed one that covers the specific nature of clinical language. Thus, we decided to use the dmis-lab/BioBERT model[6] that covers the distributional patterns of biomedical language. Unfortunately, there is no specific German clinical or biomedical language model available, yet. Thus, we used the dbmdz/bert-base-german-cased model[7] for fine-tuning. Both language models are limited to a sequence length of 512

[3] https://github.com/dmis-lab/biobert-pytorch
[4] https://huggingface.co/
[5] https://github.com/microsoft/nni
[6] https://huggingface.co/dmis-lab/biobert-base-cased-v1.2
[7] https://huggingface.co/bert-base-german-cased

tokens. Unfortunately, this limit does not account for the length of all documents in the datasets. The documents needed to be cropped with the least possible information loss. Hence, decided to center the relation-specific entities by calculating the token distance between the entities and setting the window margins at half of the remaining token distance for both sides. We also varied the token distances to measure the effect of context on our experimental results.

3. Results

Next, we will present our classification results on all three datasets, as well as the performance when varying the context window available to the model. There are two classes in the German dataset, namely *Mode* and *Reason*, and one in the I2B2 dataset, Duration, where the classification results deviate from the other two rather similar results in the other datasets. Especially for the *Reason* class this is not a surprise. This class was already reported to suffer from the lowest inter-annotator agreement values, compared to all other classes [7]. But all in all, the BERT models show comparable and stable results on all of the three datasets and both languages.

Table 1. Average macro F1-scores of the medication entity and relation extraction experiments, with Precision and Recall in brackets.

Task	Token distance	I2B2 2009	N2C2 2018	3000PA$_J$
Entity extraction	512	0.88 (0.87/0.89)	0.93 (0.93/0.93)	0.84 (0.83/0.86)
Relation extraction	10	0.99 (0.99/0.99)	0.99 (0.99/0.99)	0.99 (0.99/0.99)
	25	0.99 (0.99/0.99)	0.98 (0.98/0.99)	0.87 (0.87/0.87)
	512	0.98 (0.98/0.98)	0.81 (0.85/0.79)	0.66 (0.66/0.68)

4. Discussion

As can be seen from Table 1, the larger the data set the larger the impact of context on the overall performance of relation extraction, at least in the case of medical relation extraction. Due to the nature of the relation extraction problem where the entity pair directly defines the relation type, e.g., Drug and Duration are linked using the relation Duration-Drug, we would assume to achieve F1-scores close to 1. As the BERT models performed best, and within expectations, on the smallest dataset (I2B2 2009), the performance drops for the larger English dataset (N2C2 2018) and performs worst on the largest dataset (3000PA$_J$). Thus, the additional context information provided by the larger datasets and the large text span seem to have a negative effect on the classification results. One may speculate that the extra text added too much noise.

This behavior can be interpreted in the following way: the larger the dataset and the number of samples, the more careful the sequence length should be chosen as more and more noise may be injected for model building. Interestingly, this observation also indicates that this noise injection cannot be handled by the hyperparameter maximum sequence length alone, since longer sequences will be cut only on the right-hand side and not symmetrically.

5. Conclusions

The automatic extraction of medication information and its associated attributes from clinical texts is an important task in medical NLP. In this paper, we have shown that transformer models achieve reasonable performance figures for drugs and drug-related entities, as well as the relations between those entities. This was demonstrated for two English challenge datasets, namely I2B2 2009 and N2C2 2018, as well as one German dataset, the 3000PA$_J$ corpus.

Even though we used a more general language model for the German data for fine-tuning and classification, the final results on all three datasets are very much alike in terms of overall performance. However, we think that a language model trained on a sufficient amount of real clinical data would be beneficial for both the English- and the German-speaking communities.

An issue we did not tackle is the provision of an end-to-end system. In the future, we want to combine the entity and relation classification modules by automatically detecting promising entity pairs. There are also some more challenging clinical entity and relation tasks that might benefit from transformer models. As we have shown in this paper, this is not limited to the English-speaking (clinical) NLP community.

Acknowledgements

This work was supported by BMBF within the SMITH and DIFUTURE projects under grant 01ZZ1803G and 01ZZ2009, respectively. We thank Prof. Dr. A. Scherag, Dr. D. Ammon, and all members of the Data Integration Center of the Jena University Hospital for their support. The usage of 3000PA$_J$ is based on the approval by the local ethics committee (4639-12/15) and the data protection officer of the Jena University Hospital.

References

[1] Uzuner O, Solti I, Cadag E. Extracting medication information from clinical text. J Am Med Inform Assoc. 2010 Sep-Oct;17(5):514-8, doi: 10.1136/jamia.2010.003947.

[2] Henry S, Buchan K, Filannino M, Stubbs A, Uzuner O. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. J Am Med Inform Assoc. 2020 Jan;27(1):3-12, doi: 10.1093/jamia/ocz166.

[3] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. Advances in Neural Information Processing Systems 30 – NIPS 2017. Proceedings of the 31st Annual Conference on Neural Information Processing Systems. 2017. p. 5998–6008, doi: 10.48550/arXiv.1706.03762.

[4] Li Y, Wehbe RM, Ahmad FS, Wang H, Luo Y. A comparative study of pretrained language models for long clinical text. J Am Med Inform Assoc. 2023 Feb; 30(2):340-7, doi: 10.1093/jamia/ocac225.

[5] Devlin J, Chang MW, Lee K, Toutanova K. 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, Volume 1. 2019. p. 4171-86, doi: 10.18653/v1/n19-1423.

[6] Uzuner O, Solti I, Xia F, Cadag E. Community annotation experiment for ground truth generation for the i2b2 medication challenge. J Am Med Inform Assoc. 2010 Sep-Oct;17(5):519-23.,doi: 10.1136/jamia.2010.004200.

[7] Hahn U, Matthies F, Lohr C, Löffler M. 3000PA-towards a national reference corpus of german clinical language. Stud Health Technol Inform. 2018;247:26-30.