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

© 2024 International Medical Informatics Association (IMIA) and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/SHTI231052

Detection of Medication Mentions and Medication Change Events in Clinical Notes Using Transformer-Based Models

Yuting Guo[a,1], Yao Ge[a] and Abeed Sarker[a]
[a]*Department of Biomedical Informatics, School of Medicine, Emory University, Atlanta, United States*
ORCiD ID: Yuting Guo https://orcid.org/0000-0002-8919-0888, Abeed Sarker https://orcid.org/0000-0001-7358-544X

Abstract: In this paper, we address the related tasks of medication extraction, event classification, and context classification from clinical text. The data for the tasks were obtained from the National Natural Language Processing (NLP) Clinical Challenges (n2c2) Track 1. We developed a named entity recognition (NER) model based on BioClinicalBERT and applied a dictionary-based fuzzy matching mechanism to identify the medication mentions in clinical notes. We developed a unified model architecture for event classification and context classification. The model used two pre-trained models—BioClinicalBERT and RoBERTa to predict the class, separately. Additionally, we applied an ensemble mechanism to combine the predictions of BioClinicalBERT and RoBERTa. For event classification, our best model achieved 0.926 micro-averaged F_1-score, 5% higher than the baseline model. The shared task released the data in different stages during the evaluation phase. Our system consistently ranked among the top 10 for Releases 1 and 2.

Keywords. Deep learning, text classification, named entity recognition

1. Introduction

Clinical notes contain important health-related information, and in recent years, they have been increasingly leveraged via natural language processing (NLP) methods for clinical knowledge discovery and decision-making. Mining knowledge from clinical notes can however be challenging since they often contain domain-specific terminologies, non-standard abbreviations, and misspellings [1,2]. An additional challenge when developing clinical NLP systems is access to data. Clinical notes often contain protected health information (PHI), and they are typically not shared across institutions. To promote community-driven development of NLP methods for clinical texts, there are currently several efforts in place, including the shared tasks of the National NLP Clinical Challenges (n2c2). We participated in one of the tracks of the n2c2 shared tasks, specifically, *Track 1: Contextualized Medication Event Extraction CMED*, organized by Mahajan et al. [3]. The shared task included three subtasks:

[1] Corresponding Author: Yuting Guo, 101 Woodruff Circle, 4th Floor East, Atlanta, GA 30322, email: yuting.guo@emory.edu.

medication extraction—a named entity recognition (NER) task, event classification (Event), and context classification (Context). Medication extraction aims to identify medication mentions in the clinical notes, event classification aims to detect whether the presence of a medication change is discussed (i.e., the event is the change of medication), and context classification aims to identify the clinical context for a given event along 5 dimensions: *Action, Negation, Temporality, Certainty,* and *Actor*. The dataset consists of 9,012 annotated medication mentions in 500 clinical notes, 400 of which were available for training, and 100 were held out for evaluation. In the evaluation phase, the test data were released in different stages, and our team participated in Release 1 and 2. The data of Release 1 consist of clinical notes without any annotation, which aims to evaluate NER, NER+Event, and Context tasks (E2E). The data of Release 2 consist of the clinical notes with gold-standard medication, which aims to evaluate Event and Event+Context.The data of Release 1 and 2 both contain 100 samples.

Our team developed a NER model using BioClinicalBERT [4] and applied a dictionary-based fuzzy matching mechanism to identify possible medication mention candidates in the provided clinical notes. We also developed a unified model architecture for event classification and context classification using BioClinicalBERT and RoBERTa [5], and applied an ensemble mechanism for making the final decision. Our best model achieved 0.93 micro-averaged F_1-score, which is 0.05 higher than the baseline model [3] for event classification. Our team consistently ranked as one of the top 10 teams for Releases 1 and 2, out of between 76 and 34 teams, depending on the task. We released our code on Github: https://github.com/yguo0102/n2c2_2022_classification.

2. Methods

2.1. Data Preprocessing

The original data released by the organizers included 350 clinical notes in the training set and 50 in the development set. We re-split the 400 clinical notes into 3 data sets—300 for training, 20 for development and optimizing our models, and 80 for testing performance. This was the same distribution as the baseline system [3]. For NER, we split each clinical note into sentences and ran the model on each sentence to detect any medication mention. For event classification and context classification, we extracted the sentences around the medication mentions as the text input for the model. To avoid including irrelevant notes, we limited the length of the text input to about 100 words including 50 words before and after the medication mention, respectively. We also replaced the specific medication mentions with a normalization term <*MED*> in an attempt to reduce the noise for the classification model.

2.2. NER Model

Following standard practice, we treated the NER task as a tagging task and developed a NER model based on BioClinicalBERT. We chose BioClinicalBERT because the model was pre-trained on clinical text which is close to the shared task data. The input sentence was encoded by BioClinicalBERT into a matrix consisting of the vector representation (i.e., embedding) of each token. The token embedding vectors were then fed into a linear layer and an output layer with softmax activation, respectively. For each token, the model output a probability vector, and the tag with the highest probability was chosen during the inference phase. Additionally, we created a medication name dictionary using the knowledge base of the Unified Medical Language System (UMLS). We applied a dictionary-based fuzzy matching mechanism to predict the medication and combined the results with the NER model. Specifically, we looked for the text pieces in the clinical notes that partially matched the entries in the medication dictionary as approximate matches. We used the Levenshtein ratio measure to compute these fuzzy matches. If the matches were not recognized by the NER model, we added the matches to the final NER results. We anticipated that adding the fuzzy matches would help boost the recall of the NER system with little loss in precision.

We used strict F_1-score and lenient F_1-score to evaluate the model performance on the test set. Specifically, the strict F_1-score considers the system extraction to be correct if the character offsets exactly match those from the manual annotation, whereas the lenient F_1-score considers any overlapping between the two offset pairs as correct.

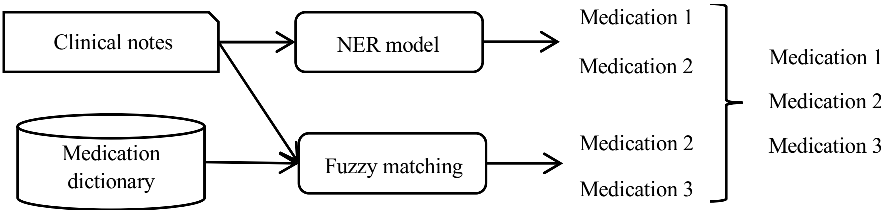

Figure 1. The framework for our medication extraction system.

2.3. Classification Model

We used the unified model architecture for event classification and context classification. The model contained a pre-trained encoder, a pooling layer, a linear layer, and an output layer with softmax activation. For each text input, the encoder encoded each token into an embedding vector, the pooling layer computed a document embedding by averaging the token embeddings, and then the document embedding was fed into the linear layer and the output layer. The output layer generated a probability vector with values between 0 and 1, and the class with the highest probability was chosen during the inference phase. For context classification, we treated the task as 5 classification tasks for 5 dimensions and independently trained the model for each dimension. We experimented with BioClinicalBERT and RoBERTa and also applied an ensemble mechanism to combine the results of the two models. Specifically, we averaged the probability vectors generated by the two models and chose the class with the highest probability.

For event classification, micro F_1-score and macro F_1-score were used as the evaluation metrics; for context classification, the combined F_1-score was used, which aims to evaluate all 5-dimensional values for a medication combined as "Combined" (e.g., *Start* + *Past* + *Certain* + *Physician* + *NotNegated*).

3. Results

Table 1 shows the performance of our systems on the dataset for Release 1. We trained three NER models with different settings. *Biocl_15* is the NER model trained for 15 epochs; *Biocl_6_dict* and *Biocl_15_dict* are the models that combined the model trained for 6 epochs and 15 epochs with the dictionary-based fuzzy matching mechanism, respectively. For event classification and context classification, we used the ensemble model on the output of the NER models. In the results, we observed that *Biocl_15_dict* consistently outperformed the other two models for NER, event classification, and context classification. Also, the dictionary-based fuzzy matching mechanism significantly improved the strict F$_1$-score but did not improve the lenient F$_1$-score for NER.

Table 1. The NER, event classification, and context classification results of Release 1.

Model	NER Strict F$_1$	NER Lenient F$_1$	Event Micro F$_1$	Event Macro F$_1$	Context Combined F$_1$
Biocl_15_dict	**0.907**	**0.972**	**0.907**	**0.817**	**0.497**
Biocl_15	0.848	**0.972**	0.882	0.769	0.479
Biocl_6_dict	**0.907**	0.968	0.886	0.771	0.475

Table 2 shows the performances obtained by our methods for the data from Release 2. For Release 2, we trained three classification models for event classification and context classification. As we can see, RoBERTa and Ensemble achieved similar performance for event classification, and RoBERTa outperformed BioClinicalBERT for context classification. For event classification, compared to the baseline which achieved 0.88 micro F$_1$-score and 0.79 macro F$_1$-score [6], all of the 3 models obtained better performance compared to context classification.

Table 2. The event classification and context classification results of Release 2.

Model	Event Micro F$_1$	Event Macro F$_1$	Context Combined F$_1$
RoBERTa	**0.926**	**0.834**	**0.544**
BioClinicalBERT	0.910	0.791	0.520
Ensemble	0.924	0.812	0.539

For Release 1, our system achieved 8th place out of 76 systems for NER, 5th place out of 52 systems for NER+Event, and 6th place out of 34 systems for E2E. For Release 2, our system achieved 10th place out of 54 systems for Event, 9th place out of 38 systems for Event+Context. Our simple approach consistently remained among the top 10 for all the tasks.

4. Discussion

For the NER model, the results show that longer training time can improve the model performance on this task. The additional dictionary-based fuzzy matching method can also help boost the performance of the pre-trained deep learning models. By error analysis, we found that the errors in the NER results can cause the errors in the classification results. A further study with more focus on building a system with independent NER and classification modules is therefore suggested. For the classification tasks, the non-domain-specific model (i.e., RoBERTa) slightly outperformed the domain-specific model (i.e., BioClinicalBERT), which suggests that domain adaptive pre-training may not improve the model performance on domain-specific tasks. This is not surprising since we have observed similar results in our prior text classification benchmarking work [7]. We also observed that the ensemble model slightly underperformed compared to RoBERTa for event classification and context classification. One possible reason is that RoBERTa and BioClinicalBERT tend to perform similarly on this task but cannot complement each other. In our future work, we will further explore more individual models and ensemble strategies to improve our performance. We also aim to explore additional pretraining strategies.

5. Conclusions

In this work, we built two Transformer-based models for the detection of medication mentions and medication event changes in clinical notes. We developed an NER model using BioClinicalBERT and applied a dictionary-based fuzzy matching mechanism to identify the medication mentions, and we developed a unified model architecture for event classification and context classification using BioClinicalBERT and RoBERTa and applied an ensemble mechanism. Our RoBERTa model achieved 92.6% micro-averaged F_1-score, 5% higher than the baseline model. For the shared task, our team is one of the top 10 teams for Release 1 and 2.

Acknowledgments

Research reported in this publication was supported by the National Institute on Drug Abuse under award number R01DA057599. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

[1] Dalianis H. Clinical text mining: Secondary use of electronic patient records. Springer Nature; 2018. doi:10.1007/978-3-319-78503-5.

[2] Spasić I, Livsey J, Keane JA, Nenadić G. Text mining of cancer-related information: Review of current status and future directions. Int J Med Inform. 2014 Sep;83:605-23, doi: 10.1016/j.ijmedinf.2014.06.009.

[3] Mahajan D, Liang JJ, Tsou C-H. Toward Understanding Clinical Context of Medication Change Events in Clinical Narratives. AMIA Annu Symp Proc. 2022;2021:833-42.

[4] Alsentzer E, Murphy JR, Boag W, Weng W-H, Jin D, Naumann T, Mcdermott MBA. Publicly Available Clinical BERT Embeddings. 2019 Apr; 1:72-78, doi: 10.48550/arXiv.1904.03323.

[5] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692. 2019 Jul, doi: 10.48550/arXiv.1907.11692.

[6] Mahajan D, Liang JJ, Tsou C-H. Toward Understanding Clinical Context of Medication Change Events in Clinical Narratives. In: AMIA Annual Symposium Proceedings. 2021; 2021: 833-42.
[7] Guo Y, Ge Y, Yang Y-C, Al-Garadi MA, Sarker A. Comparison of pretraining models and strategies for health-related social media text classification. Healthcare (Basel). 2022 Aug;10(8):1478, doi:10.3390/healthcare10081478.