Detecting Medications and Adverse Drug Events in Clinical Notes Using Recurrent Neural Networks

Xi Yang ALEXGRE@UFL.EDU
Jiang Bian BIANJIANG@UFL.EDU
Yonghui Wu [*] YONGHUI.WU@UFL.EDU
Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, Gainesville, Florida, USA

Editor: Feifan Liu, Abhyuday Jagannatha, Hong Yu

Abstract

Early detection of Adverse Drug Events (ADEs) from Electronic Health Records (EHRs) is an important, challenging task to support pharmacovigilance and drug safety surveillance. The authors present a Recurrent Neural Network (RNN)-based system to detect medication name and its attributes (dosage, frequency, route, duration), as well as mentions of ADEs, Indications, other signs and symptoms from clinical notes. We developed an RNN-based Named Entity Recognition (NER) system implemented using Long-Short Term Memory (LSTM). Two NER models, RNN-1 and RNN-2, were developed using different training strategies. Both of the two models only utilized pretrained word embeddings provided by organizers without any extra feature engineering. The RNN-2 model achieved a top 3 performance (F1-score of 0.8233) for sub-task 1, demonstrating the efficiency of RNN for clinical NER tasks.

Keywords: Clinical Natural Language Processing (NLP), Name Entity Recognition (NER), Recurrent Neural Network (RNN), LSTM

1. Introduction

Unstructured clinical text has been increasingly used for clinical and translational research as it contains detailed patient information that cannot be captured in abstracted medical codes (Meystre et al.). A well-known challenge to use unstructured clinical text for research is that much of the detailed information is documented in a narrative manner, which is not directly accessible to clinical applications. Clinical Natural Language Processing (NLP) is the key technology to extract information from unstructured clinical text to support various clinical studies and applications that depend on structured data. The clinical NLP community has organized shared tasks such as i2b2 (The Center for Informatics for Integrating Biology and the Bedside) challenges, SemEval (International Workshop on Semantic Evaluation) challenges and ShARe/CLEF eHealth Challenges, to examine the performance of current NLP methods for clinical concept and relation extraction. Different clinical Named Entity Recognition (NER) systems have been developed from these challenges based on various approaches including Conditional Random Fields (CRFs) (Lafferty et al.), Support

[*] Corresponding author.

Vector Machine (SVM) (Joachims) and Hybrid methods (deBruijn et al.; Jiang et al.). We also have explored several deep learning approaches for clinical NER (Wu et al., c,a,b).

Adverse Drug Events (ADEs) are critical for patient safety. Early detection of ADEs from EHRs provides an efficient way of pharmacovigilance and drug safety surveillance. In 2018, the University of Massachusetts Medical School organized an NLP challenge for detecting Medication and Adverse Drug Events from electronic health records (MADE1.0). We participated this challenge as the team "Gators", from the Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida. We developed a machine-learning based clinical NER system for sub-task 1 - automatically detect mentions of medication name and its attributes (dosage, frequency, route, duration), as well as mentions of ADEs, indications, other signs and symptoms. Our system achieved a performance among the top three best systems for sub-task 1. This report describes our clinical NER system developed for this challenge.

2. Methods

2.1. Datasets

In this challenge, we used a total number of 1,089 de-identified clinical notes of cancer patients provided by organizers. The clinical notes were annotated with medication information (medication name, dosage, route, frequency, duration), ADEs, indications, other signs and symptoms, and relations among those entities. The notes were divided into a training set of 876 notes for model development and a test set of 213 notes for evaluation. To facilitate the development of machine-learning models, we further divided the 876 notes into a short-training set of 776 clinical notes and a validation set of 100 clinical notes. We used the short-training set of 776 notes to train machine models and select the top-performed models according to the validation performance on develop set. Table 1 shows detailed information about the Training, Short-training, Develop and Test datasets. The study was approved by the University of Florida Institutional Review Board.

Table 1: Descriptive statistics of datasets

Dataset	**Notes**	**Entities**
Training	876	66,932
Short-training	776	60,091
Develop	100	6,841
Test	213	11,172

2.2. Name Entity Recognition

Machine-learning based NER methods typically represent the training corpus using "BIO" format and most of them identify entities on the sentence level. In this challenge, the clinical notes are provided without pre-processing of tokenization and sentence boundary. Therefore, we developed a pre-processing pipeline to normalize the raw notes (tokenization and sentence boundary) and convert the annotation from “BIOC” format into “BIO” format. Since tokenization and sentence boundary detection will change the offsets (the start and end positions) of named entities, we tracked the position mappings between the original notes and the normalized notes and recorded them using mapping files. Then, an NER module was developed to read the normalized notes and identify clinical named entities. We also developed a post-processing pipeline to read “BIO” format predictions from the NER module and generate “BIOC” format results that required for submission. The post-processing pipeline also transformed the positions of named entities back to their original positions in the raw notes. Figure 1 shows an overview of our NER system.

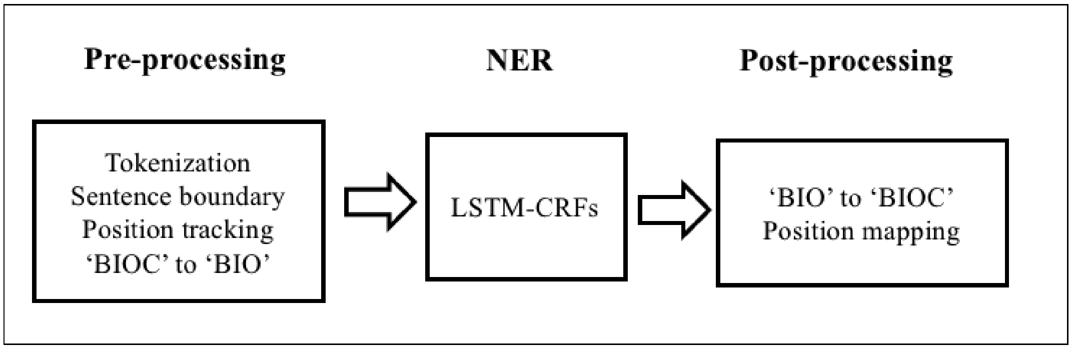

Figure 1: An overview of the NER system.

For the NER module, we applied a Recurrent Neural Network model implemented using the Long Short-Term Memory (LSTM). We adopted an LSTM-CRFs architecture from Lample et al (Lample et al.). This LSTM-CRFs composes of a character-level bidirectional LSTM layer, a word-level bidirectional LSTM layer and a CRFs layer.

2.3. Experiments, Submissions and Evaluation

We used the short-training set of 766 notes to train an LSTM-CRFs model and optimize the parameters according to the performance on the validation set of 100 notes. Pre-trained word embeddings provided by organizers (Jagannatha and Yu, a,b) were used to initiate the word embedding layer. The character embeddings were randomly initiated at the beginning and later updated during training. The optimized LSTM-CRFs model according to the validation corpus used the following parameters: the word embedding dimension was 200; the character embedding dimension was 25, the bidirectional word-level LSTM had an output dimension of 200; and the bidirectional character-level LSTM had an output size of 25; the learning rate was fixed at 0.005; the input layer for the word-level LSTM applied a dropout at probability of 0.5; the stochastic gradient descending applied a gradient clapping at [-5.0, 5.0]. We developed the system in Python under version 2.7.14.

We submitted two runs for sub-task 1, RNN-1 and RNN-2. RNN-1 is the best model trained on the short training set according to the performance on the validation set. For RNN-2, we combined the short training set and the validation set and re-trained the LSTM-CRFs model according to the parameters optimized in RNN-1. The official evaluation script provided by the challenge organizers was used to calculate the micro-average precision, recall and F-1 score (strict vs relaxed) for evaluation.

3. Results and Discussion

Table 2 shows the best performance of RNN-1 on the validation set of 100 notes. The best RNN-1 model achieved an F1-score of 0.8897 overall entities. We were unable to evaluate RNN-2 since the validation set was merged with training.

Table 2: Best performance of RNN-1 on the validation set.

Model	Performance		
	Percision	Recall	F1-score
RNN-1	**0.8893**	**0.8900**	**0.8897**

Table 3 shows the performances (micro-average strict and relaxed scores) of RNN-1 and RNN-2 on the test set. RNN-2 outperformed RNN-1 with the best strict F1-score of 0.8233. Compared with the best performance of RNN-1 on the validation set (0.8897), the best performance of RNN-2 on the test set dropped.

Table 3: Models performances on the test set.

Model	Scoring	Performance		
		Percision	Recall	F1-score
RNN-1	Strict	0.8034	0.8236	0.8134
RNN-2	Strict	**0.8149**	**0.8318**	**0.8233**
RNN-1	relaxed	**0.8214**	0.8549	0.8378
RNN-2	relaxed	0.8180	**0.8654**	**0.8411**

Table 4 shows the performance of RNN-2 on test data for each entity type. While RNN-2 achieved good performance for most of the entity types, the performance for Indication and ADE entities are notably lower than other categories.

We analyzed errors in the RNN-2 model. Some false negatives were caused by boundary mismatch and misclassification of semantic types. For example, a common type of error for ADE and Indication is to misclassify them as SSLIF. This may be caused by the limited number of training samples (ADE and Indication only accounted for about 2% and 5% of the total number of entities, respectively). Some entities were annotated with two different semantic categories and our RNN-2 could not handle them correctly. For example, the entity “attention and concentration span has decreased” is annotated both as ADE and SSLIF. There were also complex entities such as “nodular-sclerosing stage IIa Hodgkin disease”, which was annotated as Indication and part of it, “stage IIa”, was annotated as Severity. The RNN-2 was able to extract entities “nodular-sclerosing” as SSLIF, “stage IIa” as Severity, and “Hodgkin disease” as Indication but failed to detect the whole sequence

Table 4: Performance of RNN-2 on the test data for each entity category.

RNN-2	Performance		
Entity Category	Percision	Recall	F1-score
Drug	0.8597	0.9003	0.8795
Indication	0.5142	0.7605	**0.6135**
Frequency	0.8467	0.8638	0.8552
Severity	0.7509	0.7832	0.7667
Dose	0.8815	0.8393	0.8592
Duration	0.6466	0.7890	0.7107
Route	0.8869	0.9274	0.9067
ADE	0.5104	0.7432	**0.6052**
SSLIF	0.8468	0.8319	0.8232

as an Indication. A possible way to further improve our system is to integrate medical knowledge with corpus-based word embeddings and design rules to handle complex entities, which is our next focus.

4. Conclusion

This study presents our clinical NER system developed in the MADE1.0 open challenge. Our system is among the top 3 ranked systems with a final F1-score of 0.8233, demonstrating the efficiency of deep learning models for clinical NER.

References

Berry deBruijn, Colin Cherry, Svetlana Kiritchenko, Joel Martin, and Xiaodan Zhu. NRC at i2b2: one challenge, three practical tasks, nine statistical systems, hundreds of clinical records, millions of useful features. In *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data. Boston, MA, USA: i2b2.*

Abhyuday Jagannatha and Hong Yu. Bidirectional recurrent neural networks for medical event detection in electronic health records. abs/1606.07953, a. URL http://arxiv.org/abs/1606.07953.

Abhyuday Jagannatha and Hong Yu. Structured prediction models for RNN based sequence labeling in clinical text. abs/1608.00612, b. URL http://arxiv.org/abs/1608.00612.

Min Jiang, Yukun Chen, Mei Liu, S. Trent Rosenbloom, Subramani Mani, Joshua C. Denny, and Hua Xu. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. 18(5):601–606. ISSN 1067-5027 1527-974X. doi: 10.1136/amiajnl-2011-000163.

Thorsten Joachims. Advances in kernel methods. pages 169–184. MIT Press. ISBN 0-262-19416-3. URL http://dl.acm.org/citation.cfm?id=299094.299104.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289. Morgan Kaufmann Publishers Inc. ISBN 1-55860-778-1. URL http://dl.acm.org/citation.cfm?id=645530.655813.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. abs/1603.01360. URL http://arxiv.org/abs/1603.01360.

S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, and J. F. Hurdle. Extracting information from textual documents in the electronic health record: a review of recent research. pages 128–144. ISSN 0943-4747 0943-4747.

Yonghui Wu, Min Jiang, Jianbo Lei, and Hua Xu. Named entity recognition in chinese clinical text using deep neural network. 216:624–628, a. ISSN 0926-9630.

Yonghui Wu, Min Jiang, Jun Xu, Degui Zhi, and Hua Xu. Clinical named entity recognition using deep learning models. In *AMIA Annu Symp Proc*, b.

Yonghui Wu, Jun Xu, Min Jiang, Yaoyun Zhang, and Hua Xu. A study of neural word embeddings for named entity recognition in clinical text. 2015:1326–1333, c. ISSN 1942-597X.