Proceedings of Machine Learning Research 90:25–30, 2018 Medication and Adverse Drug Event Detection Workshop

Clinical NER and Relation Extraction using Bi-Char-LSTMs and Random Forest Classifiers

Arjun Magge AMAGGE@ASU.EDU
Department of Biomedical Informatics,
Arizona State University, Scottsdale, AZ, USA - 85259
Biodesign Center for Environmental Health Engineering,
Arizona State University, Tempe, AZ, USA - 85281

Matthew Scotch MATTHEW.SCOTCH@ASU.EDU
Department of Biomedical Informatics,
Arizona State University, Scottsdale, AZ, USA - 85259
Biodesign Center for Environmental Health Engineering,
Arizona State University, Tempe, AZ, USA - 85281

Graciela Gonzalez-Hernandez GRAGON@PENNMEDICINE.UPENN.EDU
Department of Biostatistics, Epidemiology, and Informatics,
The Perelman School of Medicine,
University of Pennsylvania, Philadelphia, PA, USA - 19104

Editor: Feifan Liu, Abhyuday Jagannatha, Hong Yu

Abstract

Identifying named entities from electronic health record notes and extracting relations between the entities is a crucial task for many applications in clinical and public health informatics. In this work, we present an natural language processing pipeline consisting of a named entity recognizer for identifying 9 medical named entities in clinical notes and a random forests classifier for extracting 7 types of relations between the entities.

Keywords: named entity recognition, relation extraction, natural language processing, information extraction, machine learning, deep learning.

1. Introduction

Processing the unstructured portions, i.e. free text from electronic health records (EHRs) to extract medical entities and relationships has applications in EHR phenotyping, EHR summarization, pharmacovigilance, drug-drug interaction (DDI) studies, detecting adverse drug events (ADE) and many more.

Recognizing medical entities in patients' clinical notes is an important task for extracting information and knowledge. This task is formally known as named entity recognition (NER) and is one of the first steps in natural language processing (NLP) pipelines. It is also one of the most crucial steps in the NLP pipeline as the success of subsequent steps such as entity relation extraction and entity resolution depends on its performance.

In this paper, we present an NLP pipeline for processing clinical notes and performing the NER and entity relation extraction tasks. For the NER component, we use bidirectional long short-term memory (LSTM) units coupled with a conditional random field classifier

(CRF) at the output layer. This model has been found to be very efficient for a variety of sequence tagging and chunking tasks (Reimers and Gurevych, 2017) and has been widely used in recent years across many variations (Lample et al., 2016; Ma and Hovy, 2016) including work in the biomedical domain (Jagannatha and Yu, 2016a,b; Habibi et al., 2017).

Once the entities have been recognized, we extract entity relationships in two stages. Firstly, we use a binary classifier to filter out entity pairs based on their types such that only entity pairs with possible relations between them are selected. We then use certain features extracted from the two entities and their contexts as inputs to a random forests classifier to determine the type of relationship between them.

The main components of the NER and entity relationship extraction systems are illustrated in Figure 1. The method section describes the overall architecture, system components and hyperparameters for reproducibility. The results section reports the performance of the NER and RE tasks. The final section discusses the limitations of the system and planned future work.

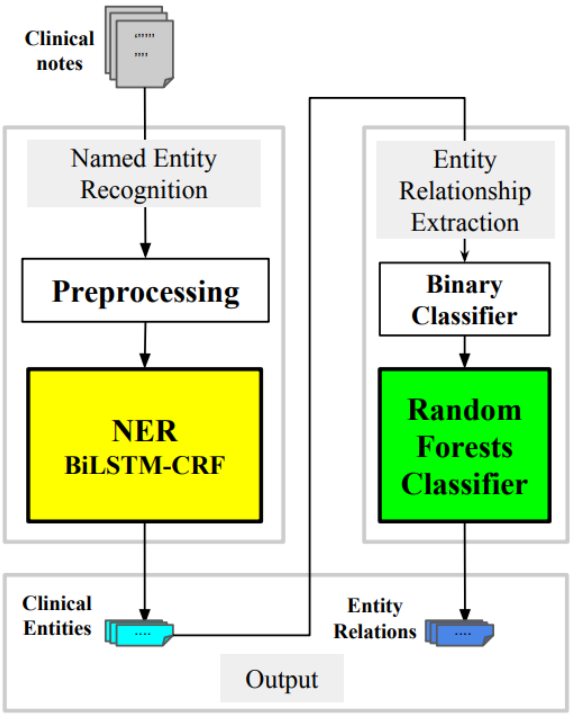

Figure 1: The overall architecture of the pipeline for the NER and Relation Extraction tasks in clinical notes.

2. Method

The gold-standard annotations for the supervised training were provided by the University of Massachusetts and contains 1092 medical notes from 21 cancer patients as part of the

MADE1.0 challenge. To train the NER, we used 800 notes as the training set, 76 as the validation set, and the remaining as the test set. (Jagannatha and Yu, 2016b,a)

2.1. Training

During preprocessing, the clinical notes are tokenized to determine sentence and token-spans. We then used the word and character embeddings of each token as inputs to train the NER as illustrated in Figure 2. We use the word embeddings developed by Jagannatha and Yu (2016a) along with character embeddings and case features. Unlike char-LSTM (Lample et al., 2016) and char-CNN (Ma and Hovy, 2016) architectures, we use a simplistic fixed size model for character embeddings. For this, we create character embeddings using the word2vec toolkit (Mikolov et al., 2013) from the training dataset with number of dimensions set to 5 and maximum number of characters set to 10. We restrict the model to use a single layer of bidirectional LSTMs, and set the number of hidden units to 75. For optimization, we use the Adam optimizer with a learning rate of 0.005 to optimize the output layer and LSTM layer variables using mean cross entropy at the output layer as loss, and CRF layer using the mean negative log likelihood. During training we use a dropout of 0.5 to prevent model overfitting.

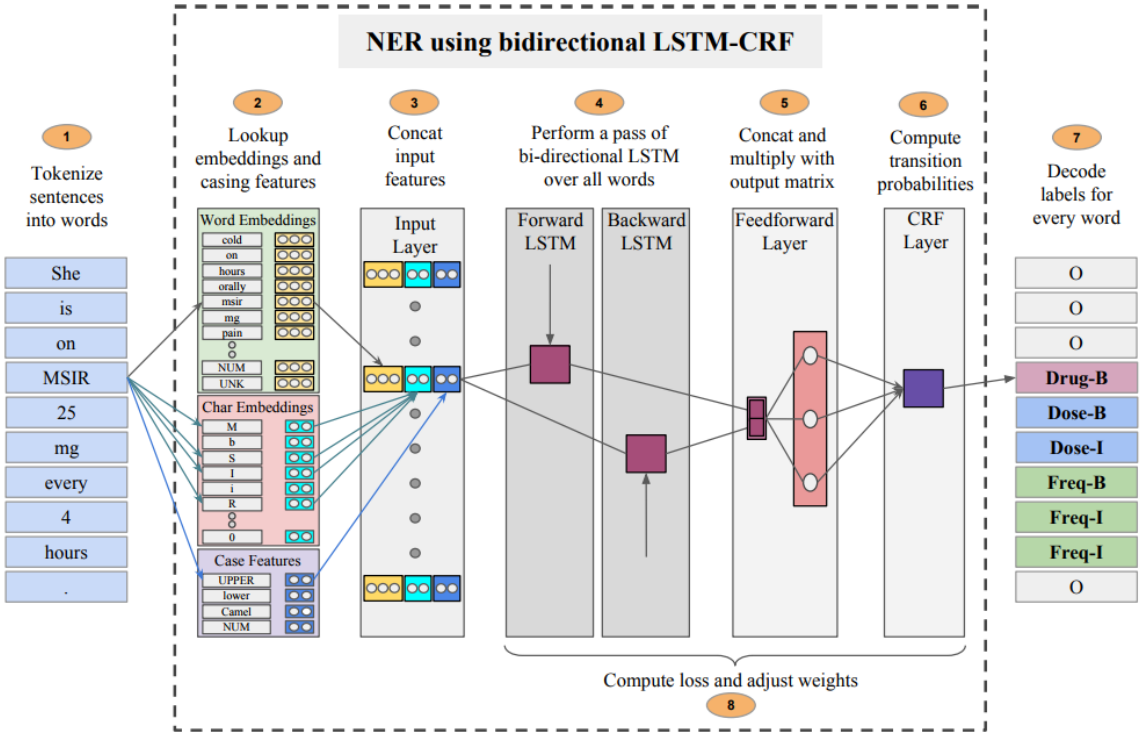

Figure 2: Steps 1 through 8 showing the training procedure of the bidirectional LSTM-CRF used for the NER task. After training has been completed only steps 1 through 7 are used to determine the labels during production.

For entity relationship extraction, since relationships between entities can exist across sentences we end up with $\binom{n}{2}$ possible relations where n is the number of named entities in the document. Hence, we first use a simple rule based binary classifier to eliminate entity pairs that cannot have any relation. We accomplish this by creating from the training set a binary distribution for the entity pairs where each value indicates if there can exist a relation or not. We then use a Random Forests classifier with 15 estimators, *gini* criterion, and minimum samples split set to 2 to classify a given input across 8 classes that includes the 7 relationship classes and 1 class for no relations. For a given pair of entities, we extract the following handcrafted features to train the classifier:

- Entity 1 type
- No. of words in Entity 1
- Avg. of entity 1 word embeddings
- Entity 2 type
- No. of words in Entity 2
- Avg. of entity 2 word embeddings
- No. of words in between entities
- Are both entities in the same sentence?

3. Results and Discussion

We created the above models using the Tensorflow and Scikit-learn libraries and used batch training to train the models. The NER presented in this paper achieved a micro-averaged F1-score of 0.82 and 0.81 during validation and testing respectively. Table 1 shows labelwise NER scores obtained on the test set. The classifier for the relation extraction task achieved an F1-score of 0.85 during validation and 0.81 during testing when gold-standard annotations were provided. In the integrated task, the system presented achieved an F1-score of 0.552 on the validation set. To obtain insight into the type of errors encountered during the relation extraction step and gather labelwise scores, we treated the first 76 clinical notes (in alphabetical order) as a secondary testing corpus. Table 1 shows labelwise scores on this secondary corpus. Due to the high disparity in performance scores on the official test corpus and the secondary corpus, we note that the the secondary corpus does not seem to be a good representative of the overall corpus. However, it does offer insight into the distribution of errors across the different entity relationship labels.

The NER's performance was found to be substantially better on medication related entities i.e. *Drug*, *Route*, *Frequency* and *Duration* compared to disorder related entities i.e. Indication, *OtherSSD*, *Severity*, and *ADE*. This difference could be attributed due to the higher number of tokens per entity in the disorder related entities. Examples of such disorder related entities that failed to be extracted include *anteverted mobile uterus without gross enlargement* and *Skin of the vulva and perineum are without lesion.* The models seemed

to achieve better precision than recall for many entities suggesting that gazetteer features might be beneficial in improving the performance of the NER and the overall system.

Similar observations could be made in the entity relationship extraction task where relation classes that involved medication entities in the same sentence were easier to classify correctly. *Dosage*, *Frequency Manner/Route* relationship classes obtained better performance than *Duration* relation where the *Duration* entity can reside on other sentences. Among disorder relation classes, *Severity* relation had significantly better classification scores on an average compared to *Reason* and *Adverse* relations where the entity pairs often reside across sentences.

Table 1: Labelwise Performance Scores for the NER task (Task 1) and the entity relationship extraction task (Task 2). The official test set was used to obtain the labelwise scores for the NER task. *For the entity relationship extraction task, The first 76 files of the training set sorted by name was heldout for obtaining labelwise test scores.

Task	**Label**	**Precision**	**Recall**	**F1-score**
Named Entity Recognition	Drug	0.87	0.84	0.86
	Dose	0.88	0.83	0.86
	Route	0.89	0.91	0.90
	Frequency	0.82	0.81	0.82
	Duration	0.74	0.82	0.78
	Indication	0.47	0.65	0.55
	Severity	0.80	0.79	0.80
	SSLIF	0.83	0.82	0.82
	ADE	0.38	0.68	0.49
	Micro-Avg	**0.80**	**0.82**	**0.81**
Relationship Extraction*	Dosage	0.88	0.94	0.91
	Manner/Route	0.93	0.97	0.95
	Frequency	0.85	0.96	0.90
	Duration	0.88	0.96	0.92
	Severity Type	0.95	0.98	0.97
	Reason	0.60	0.82	0.70
	Adverse	0.66	0.88	0.76
	Micro-Avg	**0.82**	**0.94**	**0.88**

4. Limitations and Future Work

In this work, we present an NLP pipeline to recognize 9 different medical entities in clinical notes and extract 7 possible relations in between the entities. While the system presented is more efficient and generalized than dictionary and rule based approaches, one of its limitations is its execution time which may be slower compared to the dictionary based approaches. In addition, the system presented does not yet map entities to their normalized

medical concepts. In the future, we intend to improve the performance of the system by including UMLS based gazetteer features as input to the NER. We also intend to adopt attention based convolutional neural network model for extracting relations in the between the entities.

Acknowledgments

We thank the researchers at the University of Massachusetts Lowell, Worcester, Amherst for providing the corpus used for training and testing this system. The authors gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

References

Maryam Habibi, Leon Weber, Mariana Neves, David Luis Wiegandt, and Ulf Leser. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14):i37–i48, 2017.

Abhyuday N Jagannatha and Hong Yu. Bidirectional rnn for medical event detection in electronic health records. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2016, page 473. NIH Public Access, 2016a.

Abhyuday N Jagannatha and Hong Yu. Structured prediction models for rnn based sequence labeling in clinical text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2016, page 856. NIH Public Access, 2016b.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT*, pages 260–270, 2016.

Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1064–1074, 2016.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

Nils Reimers and Iryna Gurevych. Reporting score distributions makes a difference: Performance study of lstm-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 338–348, 2017.