Title: Identifying Adverse Drug Events and Relationships in Clinical Text with Multi-task Learning

First Author:

Quan Gan, MS

Center for Data Science, New York University, (qg323@nyu.edu)

Co-Authors:

Isabel Metzger, MS

School of Medicine, New York University (isabel.metzger@nyulangone.org)

Lee Tanenbaum, MS

Center for Data Science, New York University, (lt1503@nyu.edu)

Griffin Adams, MS,

Flatiron Health, (gta@andrew.cmu.edu)

Narges Razavian, PhD

School of Medicine, New York University (narges.razavian@nyulangone.org)

Kyunghyun Cho, PhD

Center for Data Science, New York University (kyunghyun.cho@nyu.edu)

Keyword: Machine Learning, Drug-Related Side Effects and Adverse Reactions, Natural Language Processing

Abstract: Clinical narratives, such as emergency room (ER) discharge notes and physician visit notes, encode rich clinical data (such as medications, adverse events, drug administrations, and their relations) but in unstructured format. Corollary data for these data points often do not exist in structured electronic health record (EHR) data sources. Effective extraction of meaningful clinical data from EHR data, then, requires tools to automatically identify clinical entities and relationships from natural language. An example would be doctors noting *rashes* as an adverse drug event (ADE) of *beta-lactams (a DRUG)*. Detection of such entities can be formulated as a named entity recognition (NER) problem in natural language processing (NLP). For all the detected entities, the next step is to determine whether two given descriptors are related (e.g. whether rash is indeed an ADE of beta-lactams in this particular clinical note).

In this paper, we propose a multi-task learning model which tries to detect all the descriptors and their relationships in an end-to-end manner on paragraphs/sections, separated by blank lines. We formulate the NER task as predicting a sequence of begin-inside-outside (BIO) tags. The foundation of the NER task is a Bidirectional LSTM followed by a trainable conditional random field (CRF). We use pre-trained word embeddings which we learn with FastText on a custom domain-specific medical corpora. These sources include web-scraped Federal Drug Administration drug recalls, black box warning labels, synthetic clinical notes from sources such as pharmacy textbooks, and others to deal with out of vocabulary words (common in clinical corpora due to typos, abbreviations, and generic medication names). We augment the word level embeddings with character-level embeddings for which we use a Convolutional Neural Network (CNN) to synthesize. We train a binary classifier for relationship detection, which accepts as input the average of the LSTM hidden states of the corresponding words in given two phrases. We use negative sampling to balance the training procedure. We train the model by minimizing a weighted sum of the negative log-likelihood of the NER model and the ranking loss of the binary classifier. We took 38 notes from the official training set as validation set, and trained the model on the rest 265 notes (with 9374 paragraphs in total). The resulting NER model has achieved an overall lenient micro F1 score of 0.900 when evaluated by the official script on our validation set.