Entity Recognition and Disambiguation for Adverse Pharmaceutical Reactions

Rohan Kumar

Boston University

roku@bu.edu

Introduction

Entity linking, also referred to as named-entity recognition and disambiguation (NERD), is the natural language processing task of recognizing and disambiguating named entities to a knowledge base. The target knowledge base is a store of complex structured data (i.e. an ontology). Rule-based methods struggle because of ambiguous terms that can appear similar to several items in the ontology.

Motivation

In unstructured medical data, such as drug labels, there can be instances of ambiguous references to diseases, conditions, or medications. We apply several approaches to systematically identify adverse reactions in pharmaceutical drug notes/labels, and then disambiguate these reactions to a standardized medical ontology. Entities are diseases, acute conditions, and more general side effects. An implementation of this system could improve medical record administration by providing a tool to accurately connect unstructured drug reports with conditions, improving data quality and consistency.

Data and Ontology

The dataset used for both tasks is the Text Analysis Conference 2017 Adverse Drug Reaction Extraction dataset (TAC-ADR-2017), which is a collection of annotated drug labels. For each drug in the train and test set, the data includes the following:

Text: sourced from warnings, packaging labels, and general information

Entities: a list of known entities within the texts, as well as the entity class

Reactions: a list of disambiguated adverse reactions, and corresponding MedDRA ID

For NER, we convert text to sentences with BIO tagging. The MedDRA knowledge base consists of 85,392 clinically validated conditions, each with a short description and unique ID.

Research Objectives

Applying several recognition and disambiguation models to the TAC-ADR-2017 dataset, there are two main tasks we hope to perform for each drug and their corresponding text data:

(1) **Entity recognition** across the following entity classes: `AdverseReaction`, `Animal`, `Severity`, `Factor`, `Negation`, `DrugClass`

(2) **Entity disambiguation** from mentioned adverse reactions to MedDRA IDs

Models: Task 1

We utilize the BiLSTM-CRF model. BiLSTM (bidirectional long-short term memory) computes hidden states $\mathbf{h}_t = [\overrightarrow{\mathbf{h}}_t; \overleftarrow{\mathbf{h}}_t]$, where $\overrightarrow{\mathbf{h}}_t$ and $\overleftarrow{\mathbf{h}}_t$ are forward/backward hidden states, capturing context bidirectionally. These states are used to compute emission scores $\mathbf{s}_t$ for each class label. The CRF (Conditional Random Field) layer computes the sequence probability $P(\mathbf{y}|\mathbf{x}) \propto \exp\left(\Sigma_t(\mathbf{s}_{t,y_t} + T_{y_{t-1},y_t})\right)$, where T_{y_{t-1},y_t} are transition scores, and maximizes the log-probability of the correct label sequence via the Viterbi algorithm.

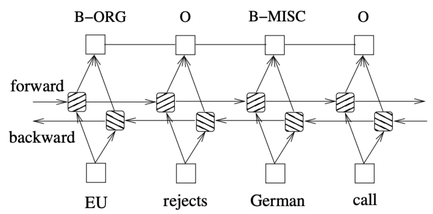

Figure 1: BiLSTM-CRF architecture

Currently, we have extensively tuned two BiLSTM-CRF models to perform entity recognition on each sentence of the drug text data:

(1) **BiLSTM-CRF; linear embeddings**

(2) **BiLSTM-CRF; BioSentVec embeddings**

BioSentVec is a word2vec-based embedding model that computes 200-dim embeddings for any given word, trained on PubMed and other medical publications.

Models: Task 2

Disambiguation can be solved via several different unsupervised approaches. We try the following techniques:

(1) **Edit Distance**: finding the minimum Levenshtein distance between the entity and an entry in MedDRA

(2) **BioSentVec Embedding**: finding most similar MedDRA entry embeddings to embedding of entity (via cosine similarity)

(3) **BioSentVec Embedding w/ Rules**: apply preprocessing before embedding such as sorting tokens and standardizing case

Experiments

For entity recognition, we evaluate our models on test data consisting of 100 drugs and their corresponding information, comparing to their known BIO tags. We utilize F1, as well as total accuracy, and accuracy excluding non-entity tags as metrics. The models were trained for 5 epochs each using the AdamW optimizer, with a hidden state dim of 128.

For entity disambiguation, we evaluate our models on all data (every drug), and compare to their known, MedDRA-disambiguated adverse reactions. Performance is measured via average F1 score, precision, and recall. In this case, F1 is a measure of how well our set of predicted MedDRA IDs overlap with the set of known MedDRA IDs.

Results

Approach (BiLSTM-CRF)	F1	Accuracy	Accuracy (non-O)
Linear embedding	0.5882	0.9328	0.6396
BSV embedding	0.7286	0.9557	0.6951

Figure 2: Task 1, recognition

Method	F1	Precision	Recall
Levenshtein	0.5496	0.4922	0.6320
BSV Embed Similarity	0.5488	0.4967	0.6200
BSV Embed Similarity w/ Rules	0.5543	0.5016	0.6261

Figure 3: Task 2, disambiguation

Discussion and Conclusion

As shown in Figure 2, the BioSentVec embedding model greatly outperformed the linear embedding. Intuitively, the linear embedding functioned only as a map between the known training vocab and embeddings, and therefore when evaluating, unknowns terms were all mapped to the same embedding, which hurt performance. BioSentVec embeddings can be generated for any word or term, and therefore we expected performance to be increased in testing. In Figure 3, we observe that all performance was similar, with the best model being the BSV embedding similarity model with preprocessing. Preprocessing by tokenizing terms and standardizing formats allows for closer embeddings, and therefore intuitively we expect better performance.

This project has been a learning experience in the difficulty of building robust NERD systems, as well as parsing highly technical medical texts. It has been clear that the use of pre-trained embedding models are crucial for achieving high performance, and we plan to expand upon this further. We plan on adding new approaches for disambiguation, and finally building an end-to-end system to convert raw medical text to disambiguated MedDRA entries.