Natural Language Processing with Clinical Notes

Rebecca Golm, Shantanu Laghate, Stephanie Wang, Jorge Ortiz, PhD
Electrical and Computer Engineering Department, Rutgers University, Piscataway, 08854

Abstract

Natural language processing (NLP) is used to help machines understand natural language. We are using NLP on clinical notes to improve named entity recognition (NER) of Protected Health Information and to identify entities such as diseases, symptoms, and medications. To develop our model, we are modifying three pretrained models – spaCy, BERT, and nltk. We train these models using the i2b2 dataset and test them on the MIMIC dataset.

Introduction

- Natural Language: the process in which humans communicate; both written and spoken[3]
- Natural Language Processing (NLP): applying machine learning to the study of language. [3]
- Named Entity Recognition (NER): identifying important information in text and categorizing it into groups. [2]
- Protected Health Information (PHI): data protected under HIPAA

Goals

- Apply NER to clinical notes to identify PHI, medications, symptoms, diagnoses, and other named entitles important in this field.

For example, in the sentence "Peter takes Zyrtec for his seasonal allergies," we want to identify Peter as the patient, seasonal allergies as a symptom, and Zyrtec as a medication. This is shown below:

Peter PATIENT takes zyrtec MEDICATION for his seasonal allergies SYMPTOM

Other NEs we want to identify are treatment, disease, allergies, and pre existing conditions.

Tools

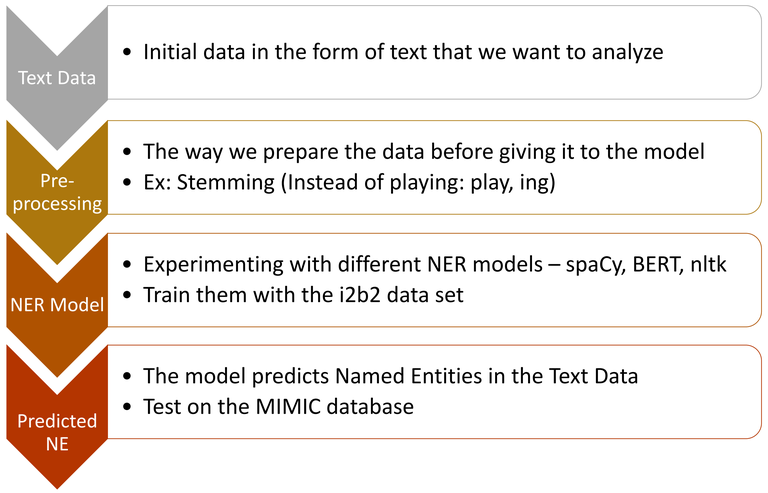
- Modifying three different NLP models - spaCy, nltk, and Bert
- i2b2: small labeled dataset (used for training)
- MIMIC: Large unlabeled dataset (used for testing)

Methods

Data Format	<PHI TYPE="Type">Text</PHI>	LABEL: …
Description	• PHI = Protected Health Information (HIPAA) • 18 Types • Ex: Dates, Phone & Fax Number • Remove to de-identify data	• Has label at sentence beginning • We can divide sentences into categories

Table 1: Format of the i2b2 dataset and uses for training NLP models

Methods



Results

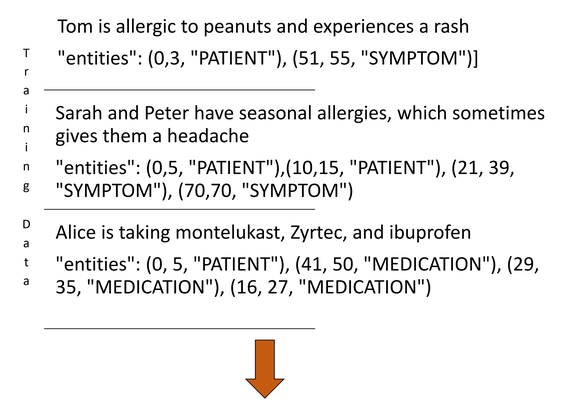


Figure 1: Result of training the spaCy Model on three sentences

Training Data: The training data consists of sentences and the entities labeled in each sentence. The entities are of the following format: start index, end index, entity name. Using the first sentence as an example, Tom is identified as a patient, since it starts a index 0 of the sentence and ends at index 3.

Results: Outcome of running the trained model on four test sentences. The output is shown with displaCy; the test data is displayed with boxes highlighting the predicted Named Entities.

Discussion

- The data above is made up as an example, since the data we use is confidential
- As can be seen, the model correctly and incorrectly identifies the NE
- In the given training data patient is always at the beginning of the sentence making the model believe that "On" is a patient
- The real data we use has more variety in sentence structure and consists of significantly more data making this flaw much less likely

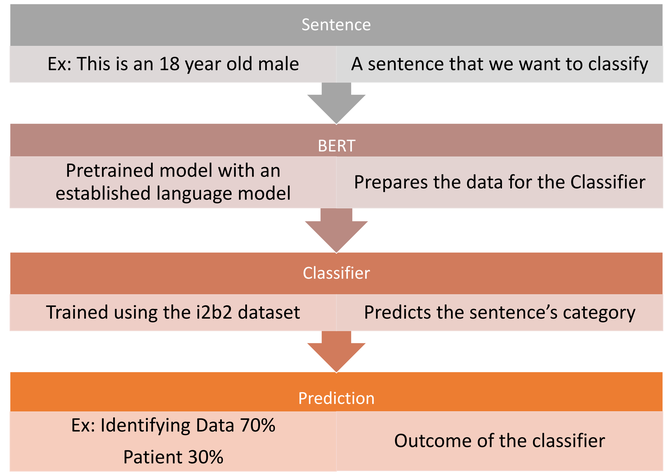
Future Directions

Currently, we have only implemented training the model with spaCy and getting more acquainted with BERT. BERT has a useful function which can tag sentences with descriptions

How BERT works:

1. Semi-supervised training on large amounts of text (Ex: Wikipedia). This was done by the developers and can be loaded in different sizes
2. Supervised training with labeled data (done by us) [1]

Process for Categorizing Sentences with BERT:



We can modify our current xml parser to look for the sentence labels and then format the data for the BERT model.

References

1. Alammar, J. (2018, December 3). The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning). Retrieved from http://jalammar.github.io/illustrated-bert/
2. Banerjee, S. (2018, November 14). Introduction to Named Entity Recognition. Retrieved from https://medium.com/explore-artificial-intelligence/introduction-to-named-entity-recognition-eda8c97c2db1
3. Brownlee, J. (2017, November 21). What Is Natural Language Processing? Retrieved from https://machinelearningmastery.com/natural-language-processing/

Acknowledgements

I want to thank Dr. Jorge Ortiz for welcoming me into his lab and for his guidance on this project. I would also like to thank everyone in the Cyber Physical Intelligence Lab for their support and mentorship. Specifically, thank you to Shantanu Laghate for helping me get started with the project and debug my code. I thank the Douglass Project and the Rutgers Electrical and Computer Engineering Department for giving me this amazing opportunity to conduct research.

RUTGERS School of Engineering

RUTGERS Douglass Residential College