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 pre-trained 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
- Natural Language Processing (NLP): applying machine learning to the study of language
- Named Entity Recognition (NER): identifying important information in text and categorizing it into groups
- Protected Health Information (PHI): data protected under HIPAA

Goals
- Apply NER to clinical notes to identify PHI, medications, symptoms, diagnoses, and other named entities 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:

![Sentence with identified entities]

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

**Text Data**
- Initial data in the form of text that we want to analyze
- The way we prepare the data before giving it to the model
- Ex: Stemming (Instead of playing: play, ing)

**NER Model**
- Experimenting with different NER models – spaCy, BERT, nltk
- Train them with the i2b2 data set

**Predicted NE**
- The model predicts Named Entities in the Text Data
- Test on the MIMIC database

Results

<table>
<thead>
<tr>
<th>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.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data: [0, 3, &quot;PATIENT&quot;], [10, 15, &quot;PATIENT&quot;], [21, 39, &quot;SYMPTOM&quot;], [41, 50, &quot;MEDICATION&quot;], [29, 35, &quot;MEDICATION&quot;]</td>
</tr>
</tbody>
</table>

**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)

Process for Categorizing Sentences with BERT:
- Ex: This is an 18 year old male
- A sentence that we want to classify
- Pre-trained model with an established language model
- Prepares the data for the Classifier
- Classifier
- Trained using the i2b2 dataset
- Predicts the sentence’s category
- Output of the classifier

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

References

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.

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

| Data Format | <PHI TYPE="Type">Text</PHI> | LABEL: ...
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>PHI = Protected Health Information (HIPAA)</td>
</tr>
<tr>
<td>18 Types</td>
<td>Ex: Dates, Phone &amp; Fax Number</td>
</tr>
<tr>
<td>Remove to de-identify data</td>
<td></td>
</tr>
<tr>
<td>Has label at sentence beginning</td>
<td></td>
</tr>
<tr>
<td>We can divide sentences into categories</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Format of the i2b2 dataset and uses for training NLP models

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

Figure 2: Result of the spaCy Model on four test sentences.