Leveraging Natural Language Processing for Automated Drug Advice

Matt Calcaterra

1 Introduction

In the realm of healthcare, accurate and efficient extraction of advice from medical texts is crucial for patient safety and informed decision-making. Specifically in the context of patients reading medically dense information, manual extraction of this information is time-consuming and prone to errors. In this project we aim to leverage natural language processing (NLP) techniques to identify advice tags for drug advice handouts and extract advice from the unstructured text. Our goal is to develop a system capable of identifying and categorizing different types of advice within drug handouts, thereby facilitating easier access to important information for healthcare professionals and patients alike.

The results of this project would specifically targeted towards future implementations of NLP models which allow for easier digestion of medical texts for patients. Ideally, with enough data, the processes taken in the project can be expanded to other medically dense texts, which once processed will become increasingly easy to digest for patient and improve patient decision-making. However, patients aren't the only ones who should be interest in the results of this project. Physician and clinical practitioners should also be interested in the results as it will provide a much quicker alternative to manually synthesizing advice from drug handouts.

2 Task Definition

The specific NLP tasks which our project will tackle will be applying named multi-class classification and sequence labeling to unstructured drug advice handouts. This project will attempt to solve the problem of unknown types of advice in drug advice handouts, and difficult medical texts for patients to parse and understand. The project will solve these problems by creating a model which takes in unstructured drug advice handouts and 1) Identifies the categories of advice in the handout and 2) identify specific portions of the text with advice and their corresponding advice tags.

What this project is not attempting to do is use drug advice handouts to determine specific contraindications, dosages, and adverse effects. Rather, the results of this project will aim to show that NLP techniques can be used on medical texts in order to make them more easily digestible by patients, and increase clarity by extracting specific advice from medically dense documents.

3 Data

Due to the confidential nature of healthcare data, annotated dataset are often difficult to acquire. For this project we will be using *A Corpus of Online Drug Usage Guideline Documents Annotated with Type of Advice*

This dataset includes two types of data: Non-Annotated Data and Annotated Data

3.1 Non-Annotated Data

The non-annotated data in this dataset consists of individual handout.txt files which contain the full text for specific drug handouts. There are **165** handouts for 165 different drugs. These texts files will be read into our code as a batch storing each handout.txt file as its own document for our model. A short segment of one of the handouts is as follows:

> *Aripiprazole is given by injection into the buttock or upper arm muscle by a health care professional, usually once every month. Some doses of some brands of this medication may also be given once every 6 weeks. Do not rub/massage the injection site after your dose.*

3.2 Annotated Data

The annotated takes the form a single tab separated value document. This document identifies specific portions of advice from the handouts, and includes annotations of advice tags. This dataset includes: Drug Name, Drug number, Line number, Advice text, Advice tags, Medication, Food, Activity, Exercise, and disease. This document only includes annotations for **90** of the non-annotated drugs.

Each drug handout has an average of **11.55** annotated advice.

4 Related Work

BioWordVec, improving biomedical word embeddings with subword information and MeSH

Offers a demonstration of word embedding in the biomedical world which differs from traditional methods computed at word level. These new methods use word vectors and embedding in order to draw contexts from the text which the word is set in. Our approach will differ in that we will be using a set of predefined labels in order to classify advice in our model.

Recent Advances in Clinical Natural Language Processing in Support of Semantic Analysis

This paper is a literature review and discussion of research from 2008 to 2014. The paper highlights three main areas of improvement for NLP in a clinical setting: more efficient corpus creation, extracting meaning from text, using NLP techniques and methods for clinical utility. The advancements and improvements found in this paper will be used to help generate a model using improved techniques.

Extracting Medical Information From Free-Text and Unstructured Patient-Generated Health Data Using Natural Language Processing Methods: Feasibility Study With Real-world Data

Attempt to determine whether an NLP pipeline is feasible for extracting medical information from free-form clinical text. Found that using a combination of NLP techniques was sufficient to create an accurate model for extracting relevant medical information. Similar to this work we will attempt to extract specific parts of interest from unstructured text, our work will differ however in that we are looking to extract and classify advice through supervised processes.

5 Evaluation

We will evaluate our model using F1-scores as out primary evaluation metric. We will calculate the precision, recall, and F1-scores using the original labels included in the annotated data in the form of advice tags. We will calculate the F1-score for each category individually in order to determine our models performance on specific categories. Additionally, we will generate a confusion matrix in order to visualize the model's performance across different classes.

Our model will then compare the results of our model to two different baselines. The first baseline will be a random performance baseline where we randomly assign the labels to our test data and assess the performance compared to our model. The second baseline we will use a simple baseline where we predict "Other drug related" which is the most common tag in the annotated dataset.

6 Work Plan

The work plan for this project will be as follows:

- Data Reprocessing: reprocess the individual handout.txt files and annotated data for training
- Labels: determine a consistent solution for label assignment
- Model Development: Design a model which classifies whole text documents. Expand the model to extract advice and then classify each extraction
- Evaluation: Evaluate our model against our two baselines

References