SI630 Project Proposal : Explaining Clinical Notes

Version 1.0

Gurpreet Singh Bhatia

Abstract

Doctors write notes in language and writing that is hard to understand. The unlegible writing can be mitigated by preparing electronic records or computer vision algorithms that converts it to a more readable format. One other problems remains, how do ordinary people understand the medical note, even after being able to read it. In this project I will work on using Natural Language processing to make more sense of such medical notes by adding more meaning to them.

1 Introduction (0.33 points)

In places like India doctors spend less than 5 minutes per patient. Sometimes advising multiple patients at a time. They cannot explain all the details to the patient in such short duration. They transcribe in a rush on a medical note but the patients cannot understand that. Such medical notes are written in a lingo not easily understood by non-doctors. Can patients atleast use the doctors note to gain more knowledge? Can we build a algorithm that explains them the nature of the medical issues? In the absence of enough medical professionals this will crucial, both helping patients and also saving doctors time.

Some organisations here at Michigan that could be interested in this project are UM Medicine (hospital).

2 Problem Definition (1 point)

Clinical notes are written in a very incomprehensible manner and can only understood by medical professionals. Looking at the text below, which is an actual doctors note for patient, we can see why. Hence patients have a very hard time drawing inference from it and have to rely on whatever the doctor spoke to them in person, which is often times very limited.

In my project my aim to is to *translate* the way doctors write to the way most ordinary humans could understand. As we can see from the example below not all portion of the text need to be explained, and the focus needs to be on hard to explain portion. Some portion of the text is just patient parameters that is also easy understood. The way I classify hard to explain would be using word probabilities of all the words and rare words will be the medical words I need to explain or *translate*
.

$DoctorsNote$:
s:a M negotiator aged 46 ys presents w/ c/o 2 days h/o mild ringing in the ears, mild headaches particularly at the back of the head and in t be he morning and mild occasional lightheadedness. o:Height 181 cm, Weight 84.7 kg, Temperature 36.9 C, Pulse 81, SystolicBP 147, DiastolicBP 94, Respiration 16, Heart = 2/6 systolic murmur at base of heart, Chest = clear to auscultation B/L, no rales or wheezing, Extremities = no edema or clubbing, Heart = normal S1, S2, RRR a:Hypertension p:performed E/M Level 3 (established patient) - Completed, and prescribed Hydrochlorothiazide - 50 mg po qd, and ordered Cholesterol.

3 Data (1 point)

Current Available data set to start with: https://data.world/arvin6/medical-records-10-yrs
This dataset contains medical records for 10 years. The table that I will be working with in *encounter.csv* and looking at *soap-notes* column. The column contains notes from doctors on the patients that I am looking to draw inference from and finally translate. One other venue I am

looking at is Open Source Vista Platform.

$EASY\_INFERENCE$:
mild headaches particularly at the back of the head and in the morning and mild occasional lightheadedness

$PARAMETER\_INFERENCE$:
o:Height 181 cm, Weight 84.7 kg, Temperature 36.9 C, Pulse 81, SystolicBP 147, DiastolicBP 94, Respiration 16, Heart = 2/6

The problem comes with:

$HARD\_INFERENCE$:
systolic murmur at base of heart, Chest = clear to auscultation B/L, no rales or wheezing, Extremities = no edema or clubbing, Heart = normal S1, S2, RRR a:Hypertension p:performed E/M Level 3

In this project I hope go beyond just fetching definitions for the medical terms. My aim is to decipher meaning from these terms and combine them to form a well founded explanation for the patient (Luo et al., 2019). I am one dataset already at hand but still looking at others just to expand my training size.

4 Related Work (0.33 points)

The idea is to merge background medical knowledge to terse doctors notes (Perera et al., 2013). One of the key challenges would to extract information using external sources as simply replacing them with definations would not solve the purpose (Narasimhan et al., 2016). Also, work to decipher meaning from ancient languages could prove really usefull for me (Luo et al., 2019).

5 Methodology

My current idea is to find crucial points for the hard part of the medical notes. then scan a medical wikipedia or webmd to extract information for them and then embed this back into the note, in a way that makes sense.
2nd idea is to do some translation on the text so that it can convert medical notes to how ordinary people speak.

6 Evaluation and Results (1 point)

One way to measure the model does is well is to see if the $\sum P(x = word_i)$ increases or not. If we are translating to a language understood by others then the words coming up should be more common and hence the $P(x = word_i)$ should ideally increase. Word probability I plan to use for checking is Wikipedia, as it offers simple extension to everyday English. This has to account for increase in sentence length.

7 Work Plan (0.33 points)

1. Gather pending datasets from Open Vista and emailing other people who do similar at Life Science Institute.

2. Write a script to fetch meaning of medical terms, using webmd or wikipedia. Dictionary of medical words that has more information than the word itself.

3. Calculate words with probabilities, using commonly available word token data. Rare words in the notes should be medical words as that is the only context here.

4. Mix medicalWord meaning with its context from the note, that is easier to understand.

5. translate to commonly understandable english – in the simple case it would require me to create to sentences from tokens, old + new. New tokens would be the word meaning for terms, replacing the word itself.

References

Jiaming Luo, Yuan Cao, and Regina Barzilay. 2019. Neural decipherment via minimum-cost flow: From ugaritic to linear B. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*. Association for Computational Linguistics, pages 3146–3155. https://doi.org/10.18653/v1/p19-1303.

Karthik Narasimhan, Adam Yala, and Regina Barzilay. 2016. Improving information extraction by acquiring external evidence with reinforcement learning. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, pages 2355–2365. https://doi.org/10.18653/v1/D16-1261.

Sujan Perera, Amit Sheth, Krishnaprasad Thirunarayan, Suhas Nair, and Neil Shah. 2013. Challenges in understanding clinical notes: Why nlp engines fall short and where background knowledge can help. In *Proceedings of the 2013 International Workshop on Data Management Analytics for Healthcare*. Association for Computing Machinery, New York, NY, USA, DARE '13, page 21–26. https://doi.org/10.1145/2512410.2512427.