Utilizing BERT pre-trained embeddings and language model for medical transcription simplification

Shannon Rumsey, Christoper Straw, Kristin Cheung
University of California, Santa Barbara
shannonrumsey@ucsb.edu

1 ABSTRACT

Using synonym replacement and text summarization, we explore the methodology of medical transcription simplification and what it means for text to be "simplified". The synonym simplification process utilizes BioBERT embeddings to identify complex terms and replace them with simpler definitions. The resulting text is then summarized using a BERT2BERT model fine-tuned on a CNN/Dailymail dataset. From the results, we use a standard method for generating reading levels, the Flesch reading ease score test, and find an overall increase in readability after applying our models.

2 INTRODUCTION

On January 20th, 2020 the first case of COVID-19 was detected in America [1] and soon after, the viral impact was considered a pandemic. Shortly after the onslaught of the virus, many were asked to confine to their homes as a means to limit the spread within the community. With many being limited socially, any understanding of news regarding the pandemic was mostly self-sustained.

Individuals were required to comprehend and act on any information given from medical outlets, which is typically written at the 6th or 7th-grade reading level [3]. This is a juxtaposition to the fact that more than half of Americans have literacy below 6th grade [5].

Those who cannot properly read or understand medical information typically "have trouble understanding medication instructions, appointment reminder forms, informed consent, discharge instructions, and health education materials" [2]. Not only does this put a strain on the individual, but "[l]ow health literacy costs the US health care system up to $73 billion annually" [2].

That being said, while there is an attempt to standardize the literacy of medical information for the sake of comprehension and readability, there is a disconnect with how it should be implemented. This is demonstrated by the gap between the current reading level of citizens and the proposed reading level for medical text. There are also very few if not no guidelines for comprehension in the field.

It is clear that the reading level of medical text should be lower than 6th grade to reflect the average American. This illuminates the need to define the 6th-grade literacy level and how we can recreate it.

We will be primarily focusing on two literacy aspects: readability and comprehension. A common indicator for readability is the Flesch reading ease score test, or FRES. This test takes into account the total number of words, number of sentences, and number of syllables in a document. From the FRES score, we can calculate the reading level grade. The FRES does not take into account the comprehension of the material, therefore, while we may be able to quantify the readability of a document, there is no concrete way to define comprehensibility.

$$206.835 - 1.015\left(\frac{\text{total words}}{\text{total sentences}}\right) - 84.6\left(\frac{\text{total syllables}}{\text{total words}}\right)$$

Figure 1: The Flesch-Kincaid score test for calculating the readability of a text [4]

Considering the variables that the FRES test is composed of, a logical next step is to take long and complicated words, or those with many syllables, and expand them into a string of shorter words that encapsulate the same information. This is the reasoning behind our synonym simplification method, we will be replacing medical terms with their definitions. To condense the longer sentences that were created by swapping terms with their definition, we then summarize using a computational linguistics model.

Our future work consists of further defining comprehensibility and its relationship to readability.

3 DATA

The dataset we used was a combination of two other datasets that had been combined for the purposes of this project. The first source for our dataset consisted of transcription examples scraped from mtsamples.com and published onto the Kaggle website. There were 2,358 unique transcriptions which we wanted to use as a medium for simplification.

The second dataset was the Harvard Health Medical Dictionary and was scraped from the Harvard Health Publishing website for this project. This file contained relatively simple medical definitions for terms we were looking to replace using synonym replacement in the transcriptions. As part of preprocessing, we generated plural forms for all nouns to account for their appearance in the transcript ("allergy" → "allergy","allergies")

4 METHODS

We will be breaking down our methodology into two parts, the synonym simplification, and the summarization.

4.1 BioBERT Embeddings

For the first portion of our methodology, we used synonym swapping for easy simplification. We preprocess the transcript for simplification by manually swapping terms with their simpler definitions. To analyze the embeddings of the terms, we used BioBERT-1.1, a modified version of BERT that was trained on PubMed data and created by DMIS-lab to specifically focus on medical terminology. First, we concatenated the term and definition, then generated context-dependent embeddings for each of the terms in the dictionary given the definitions as context and saved the term-embedding pairs to a

SUBJECTIVE:, This 23-year-old white female presents with complaint of allergies. She used to have allergies when she lived in Seattle but she thinks they are worse here. In the past, she has tried Claritin, and Zyrtec. Both worked for short time but then seemed to lose effectiveness. She has used Allegra also. She used that last summer and she began using it again two weeks ago. It does not appear to be working very well. She has used over-the-counter sprays but no prescription nasal sprays. She does have asthma but doest not require daily medication for this and does not think it is flaring up.,MEDICATIONS: , Her only medication currently is Ortho Tri-Cyclen and the Allegra.,ALLERGIES: , She has no known medicine allergies.,OBJECTIVE:,Vitals: Weight was 130 pounds and blood pressure 124/78.,HEENT: Her throat was mildly erythematous without exudate. Nasal mucosa was erythematous and swollen. Only clear drainage was seen. TMs were clear.,Neck: Supple without adenopathy.,Lungs: Clear.,ASSESSMENT:, Allergic rhinitis.,PLAN:,1. She will try Zyrtec instead of Allegra again. Another option will be to use loratadine. She does not think she has prescription coverage so that might be cheaper.,2. Samples of Nasonex two sprays in each nostril given for three weeks. A prescription was written as well

Figure 2: An example of the transcription data.

file. Then, we find any terms in the transcript, and extract context dependent embeddings for each term.

Once we get the embeddings of the term in the transcript, we perform cosine similarity between the term from the dictionary and the term in the transcript. If the similarity is above a certain threshold (set to 0.66 in our case), we replace the term in the dictionary with its definition (which is considered our synonym). The reason we need to perform this cosine similarity check is so we make sure the term we want to replace is being used in the same way as the definition ("a will [N]" vs "he will [V]"). By generating all forms of a word during preprocessing, we can account for the different forms appearing in the transcript without losing important information to lemmatization. This concludes the text replacement preprocessing for the next BERT2BERT portion.

4.2 BERT2BERT Summarization

The second portion of our methodology is where we attempt to "smooth out" or polish our simplifications. This is our way of increasing the readability of medical text. Looking at figure 3, the text seems difficult to read. Our goal for this phase of the project is to use a pre-trained BERT2BERT model that would summarize the information. The exact pre-trained model we used was the BERT2BERT Summarization with an EncoderDecoder Framework created by Patrick von Platen and published on the Hugging Face website. The model was fine tuned on the CNN/Dailymail summarization dataset on Hugging Face and while this is not ideal data for our project goal, we believe it is still a good model given the information readily available to us. From here, we tokenize the new transcriptions that resulted from the BioBert synonym swap and feed them into the BERT2BERT model.

SUBJECTIVE:, This 23-year-old white female presents with complaint of **An immune system reaction (for example, rash, fever, sneezing, or headaches) to something that is normally harmless**. She used to have **An immune system reaction (for example, rash, fever, sneezing, or headaches) to something that is normally harmless** when she lived in Seattle but she thinks they are worse here. In the past, she has tried Claritin, and Zyrtec. Both worked for short time but then seemed to lose effectiveness. She has used Allegra also. She used that last summer and she began using it again two weeks ago. It does not appear to be working very well. She has used over-the-counter sprays but no prescription nasal sprays. She does have **A disease that inflames and narrows airways, causing wheezing, shortness of breath, coughing, and tightness in the chest** but doest not require daily medication for this and does not think it is flaring up.,MEDICATIONS: , Her only medication currently is Ortho Tri-Cyclen and the Allegra.,ALLERGIES: , She has no known medicine **An immune system reaction (for example, rash, fever, sneezing, or headaches) to something that is normally harmless.**,OBJECTIVE:,Vitals: Weight was 130 pounds and blood pressure 124/78.,HEENT: Her throat was mildly erythematous without exudate. Nasal **Tissue that lines the tube-like structures of the body such as the respiratory and gastrointestinal tracts** was erythematous and swollen. Only clear drainage was seen. TMs were clear.,Neck: Supple without adenopathy.,Lungs: Clear.,ASSESSMENT:, Allergic rhinitis.,PLAN:,1. She will try Zyrtec instead of Allegra again. Another option will be to use loratadine. She does not think she has prescription coverage so that might be cheaper.,2. Samples of Nasonex two sprays in each nostril given for three weeks. A prescription was written as well.

Figure 3: The transcription data with synonyms replaced (where the synonyms were replaced is in bold).

According to the documentation for the model, in order for the pre-trained BERT2BERT model to perform, we needed to truncate the text to be no longer than 1,024 characters in length. In addition to the pre-trained model, we also used the corresponding pre-trained tokenizer.

the 23 - year - old white female has a disease that inflames and narrows airways, causing wheezing, shortness of breath, coughing, and tightness in the chest but does not require regular nasal sprays. she has used allegra, or ortho tri - cyclen, and alleges that she is allergic to claritin, and doesn't need daily medication.

Figure 4: The resulting transcription after applying the BERT2BERT model.

5 DISCUSSION

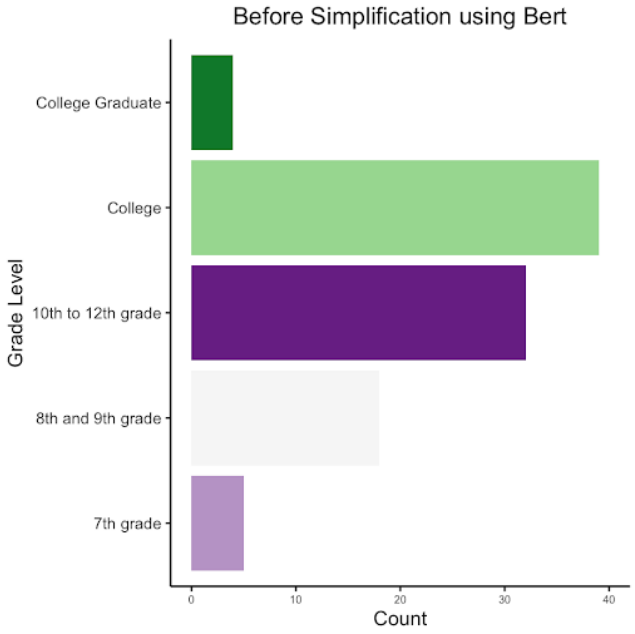

Figure 5: Reading level of the original text.

In order to determine the results of our models, we calculate the FRES values for both the text before and after our process. For illustrative purposes, we converted FRES values into grade levels. Figure 5 shows the reading level distribution of the original text favors "College" and "10th to 12th grade" readability. The medical transcriptions do not feature any examples of text that are below the 6th grade reading level, the level at which most Americans are at. This further illustrates the need for simplification of medical text.

After switching out terms with definitions based on Bert embeddings and summarizing over the text with the BERT2BERT pre-trained model, we have much better results. Figure 6 shows an overall downwards trend for the reading levels. While we expect to see results that were below the 6th grade reading level, it is still pleasant to see the model produce sound results. Instead of having the majority of text under the category "College" and "10th to 12th grade", like seen in figure 5,

we now have the majority of text in "10th to 12th grade" and "8th and 9th grade" in figure 6. Only looking at figures 5 and 6, we could make the claim that our model produced roughly a one to two grade decrease in readability.

In order to further determine the difference between the original and simplified text FRES values, we constructed a scatter plot (figure 7). The regression line indicates that the simplified text starts out with a higher FRES value and continues on a slight slope until around the FRES values of 60. Once the values are 60, it appears that the simplified and original text produce very similar scores. Please note that a higher FRES value is associated with a lower reading level (high FRES: high readability, high FRES: low grade level).

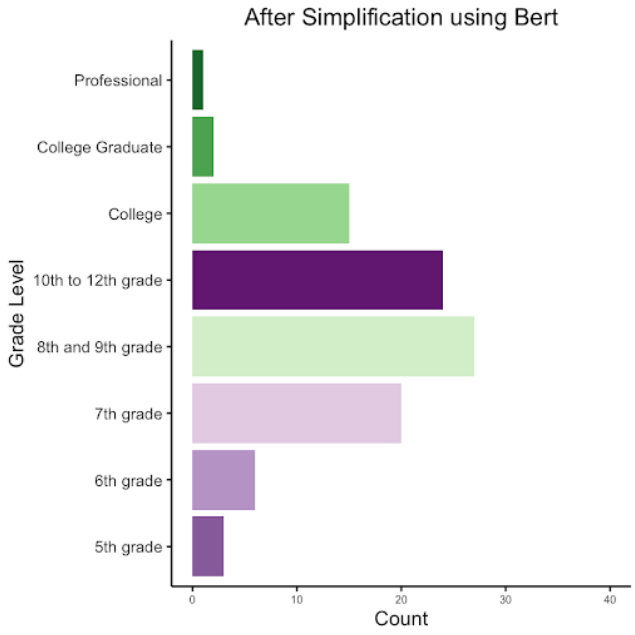

Figure 6: Reading level of simplified text.

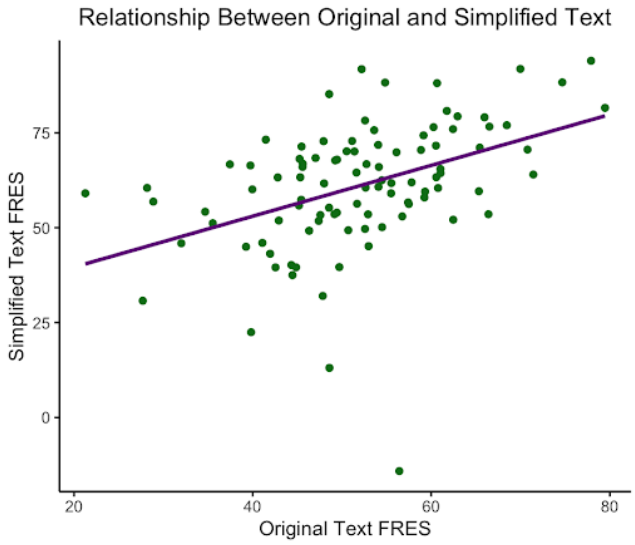

Figure 7: Scatter plot of the original text and simplified text FRES values. A linear regression model was fitted to the data to show relationships between the two variables.

5.1 Limitations and Future Work

Our model hopes to bridge the gap between medical text and consumer knowledge. With the lack of publicly available data, we were confined to models that were either pre-trained on medical data or models that were pre-trained on data from other domains. Ideally, we would extract embeddings and summarize using the same model and training data.

A potential issue with our model is how the data from the mt-samples.com website is fictional, which could make transferring this model to real data difficult or produce different results.

Another issue we may have is that the resulting summary does not convey the same information as the original text. In figure 2, the transcription claims that the woman does not think that Claritin works whereas in figure 4, the woman says she is allergic to Claritin. This could be attributed to the quality of our embeddings or our cosine similarity function.

While this project mainly addressed the readability of medical information and the potential to automate it, another goal of ours was to define and follow suit with reading comprehension as well.

A limitation of this project was the inability to hold a human focus group that could establish what it means for text to be comprehensible. If we were able to do this, the focus group would likely ask participants to fill out a questionnaire regarding their understanding of medical text and ways it could be improved upon. The results of these findings would be used as reinforcement learning from human feedback that would enable the use of human supervision to guide the creation of summaries in order to account for both readability and comprehension. Essentially, our goal is to gain human feedback as a means to fine-tune our simplification model so that it also increases the comprehensibility of text.

REFERENCES

[1] Centers for Disease Control and Prevention. [n. d.]. *CDC museum COVID-19 timeline.* Available Online: https://www.cdc.gov/museum/timeline/covid19.html.

[2] J. Graham S. Brookey. 2008. Do patients understand?. In *The Permanente Journal, 12(3).* 67–69.

[3] Baird G.L. Garg M. Hutchinson, N. 2016. Examining the reading level of internet medical information for common internal medical diagnoses. In *The American Journal of Medicine.* 129(6), 637–639.

[4] P. Von Platten. [n. d.]. . Available Online: https://huggingface.co/patrickvonplaten/bert2bert_cnn_daily_mail.

[5] Wikipedia. [n. d.]. *Literacy in the United States.* Available Online: https://en.wikipedia.org/wiki/Literacy_in_the_United_States#cite_note-2.