Medical Text Simplification

Baris Karacan
University of Illinois Chicago
Chicago, United States
bkarac3@uic.edu

Anuja Tayal
University of Illinois Chicago
Chicago, United States
atayal4@uic.edu

ABSTRACT

Thanks to rapid advancements in technology around the world, patients now have more access to health materials and they have more opportunity to engage in self-care. However, readability measures of health information is significantly higher than an average patient's health literacy abilities as most of the health documents are jargon-heavy and contain complex words that are challenging to comprehend for lay people. Due to this lack of understanding, readmission rates to hospitals are extremely high for patients who suffer critical and chronic conditions such as heart failure. To cope with this problem, we propose a medical text simplification system which utilizes two different simplification techniques: (1) replacement and (2) generation. By utilizing these techniques, we aim to create simplified versions of given medical documents. Finally, these simple documents could aid patients to increase their self-confidence and motivation, enhance their willing to engage with the treatment and reduce readmission rates of patients under critical conditions.

INTRODUCTION

Today, the web contains an abundance of useful data about wide range of medical topics[12]. Consequentially, people are now able access to these data by utilizing different social networking platforms such as Wikipedia,Reddit and Quora. Furthermore, it is also possible to obtain undisclosed electronic health records of numerous patients by completing online trainings such as CITI IRB training. Despite having access to a plethora of helpful medical information, most of the patients have difficulties of comprehending it.

Researchers from different medical disciplines pointed out that vast majority of people in the US do not have adequate health literacy to comprehend medical documents. According to [5] and [4], patients feel overwhelmed with too much information and they often have problems understanding medical terms which mostly consist of complex words. Since inadequate level of health literacy results in worse health outcomes especially for people who suffer critical conditions like health failure, there is an urgency to take actions that could close the gap between health literacy of patients and health literacy level required for the medical documents[15]. To alleviate this problem, lexical simplification could be employed to replace complex words with simpler alternatives. This simplification technique might be effective because some studies show that people who are familiar with the vocabulary of a given text could often understand its meaning regardless of the complexity of its grammatical structure[13]. Furthermore, training encoder-decoder Transformer models may also be helpful to generate simplified text from scratch. To achieve that, decoder could penalize jargon-heavy terms that are generated by the model and produce text with higher readability[6]. As a consequence, we explored two distinct approaches to simplify text in medical documents which are: (1) replacement approach and (2) generation approach.

Replacement approach is performed in 3 steps. First, complex words in each sentence of the given medical document are identified. Next, potential list of candidates are generated which are suitable for the context of the sentence. Then, the best candidates among them are determined for each complex word and replaced the context words without disrupting the syntax and semantics of the sentence. Finally, a new document is created from these sentences which consist of simpler, meaningful words that are more understandable by lay people.

But, sometimes just replacing complex words to simpler words in a sentence is not sufficient as it does not constitute a logical and meaningful sentence. The simplified sentence should be coherent, structured and follow the grammatical rules of the language. Secondly, different words have different context when used differently. So it does not make sense to replace a complex word with just a single word every time. Also, there are numerous approaches to simplify a sentence. To overcome all these limitations of Replacement approach, Natural Language Generation (NLG) is used to simplify a sentence where model is learned from scratch to generate a meaningful, coherent and grammatically structured sentence. NLG takes the help of Natural Language Understanding(NLU) Module to understand the context of the sentence, grammatical structure of the sentence and accordingly new sentence is generated. Currently, NLG techniques are used in generating responses in chatbots (Alexa,Siri etc), summarisation among others.

In sum, this work aims to enhance the medical perception of wide-range of patients who suffer from various diseases. By that way, they could be more conscious about the steps they should take to overcome their conditions. Besides, in the future, this approach could also be applied to discharge notes, and physicians could provide the simplified version of these notes to patients as a supplementary material.

ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

RELATED WORK

Recently, transformer-based NLP models have been widely used for text simplification tasks. The authors of [13] proposed a transformer-based lexical simplification framework called LSBert which is a context-aware system that detects the words to be simplified, generates substitution candidates and then ranks the candidates according to their feature qualities.

Furthermore, there were also a few works that performed text simplification in medical domain. For example, the authors of [12] proposed a denoising autoencoder framework which is trained on a dataset of simple medical sentences for simplifying the medical text. In this approach, they created a noisy corpus from simple medical sentences and trained this data by using transformer based encoder-decoder model to capture simplistic style of the sentences including grammatical structure, syntax, and choice of target words.

Moreover, an unsupervised medical text simplification system is introduced in [15]. To build this system, a 3-stage pipeline is designed. First, simple and complex words are determined with respect to their frequency of usage. Next, simplification candidates of given complex words are generated and each assigned a replacement probability. Finally, each candidate output simplification has given language model and replacement model scores, and the output with the highest total score is selected as the simple text.

In [6], the authors proposed a new metric to measure the readability and technicality of complex medical text and proposed to simplify text by generating new simplified sentences by analysing style of text and penalising the model for producing complex terms.

Text Summarization [11] is similar to Simplification where the task is to produce a concise summary while preserving the key information and the overall meaning of the document. Summarization of a document can be either extractive or abstractive. Many researchers work on summarising news articles, conversations, business meetings, customer service. Early works focused mainly on extracting important content from the documents or paraphrasing the content to form new sentences.

Past works on summarisation in medical field has been mostly on evidence based summarisation, summarising patient-nurse conversations, biomedical scientific research articles, .

BART BART [10] is sequence-to-sequence transformer model with both encoder and a decoder- bidirectional encoder (ex BERT) and a left-to-right decoder (ex GPT). It is mainly used for generation tasks but can also be used for comprehensive tasks. BART is pretrained with 5 tasks as shown in 1

- **Token Masking**- Token is masked as in BERT and is trained to predict the word
- **Token Deletion**- Token is deleted and then trained to restore it.
- **Text Infilling**- Multiple words in a span are masked with a single Mask token and is trained to predict number of words masked.
- **Sentence Permutation**- Sentences in a paragraph/context are shuffled and trained to predict the original order.
- **Document Rotation**- Document is rotated to start from random token and trained to get the start token of the document.

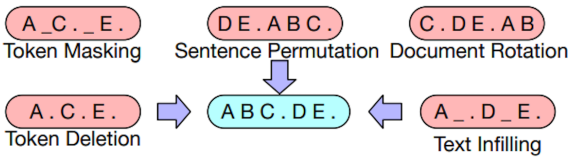


Figure 1: BART Transformation Techniques

To evaluate the simplicity of text, researchers used various readability measures also present in TextStat Python Library which relies on shallow cues of length of words, sentences and documents. These along with readability also measures complexity and grade level of the text. For example, If a text is scored a value of 9, it implies that a 9th grader will be able to understand the text. More the difference between the grade level between the complex and simple sentence, more the better. Some of the common readability scores are Automated Readability Index (ARI) [16], Flesch Kincaid Grade and smog index among others.

Decoding Strategies

For generating a sentence, effective decoding [18, 9] method plays a very important role as selecting a better model architecture does for auto regressive models. Sometimes words generated makes sense, but sometimes it starts repeating itself and text does not makes sense for which we can penalise the model.

Greedy Approach In greedy approach, next word with the highest probability is selected at each time step.

Beam Search Beam Search extends the greedy approach by choosing number of beams, and keeping most likely those subsets of hypothesis at each turn till the end.

Sampling Beam Search works well when sentence length is small but as length of the sentence increases or varies for each text, it becomes increasingly difficult to generate new meaningful sentence. Till now text generated were from a probability distribution but in [9], the authors showed that text generated is more accurate when there is an element of surprise or unpredictable. While sampling words from a random distribution to increase the probability or likelihood of high probability words and decreasing the likelihood of low probability words, we can also smooth the distribution by setting the temperature parameter. Most common sampling approaches are as follows

- **Top K Sampling**- In top k sampling k most likely words are extracted, probability mass is calculated again and then the next word is generated.
- **Top p Sampling**- As top k sampling is not able to vary the number of words filtered at each time step, top p sampling was introduced. In top p sampling [9], next word is chosen

from the set of words whose cumulative probability exceeds the probability p.

DATASET

To simplify complex medical texts, we are using Medical EW-SEW dataset which was filtered from EW-SEW dataset by QuickUMLS, a named entity recognition tool, so that in each sentence, there is atleast a mention of one medical entity of type Disease or Syndrome or Clinical Drug. In total Medical EW-SEW dataset contains 2267 pairs of complex/ordinary and simple English Wikipedia sentence pairs.

A snipet of dataset is shown in Figure 1. As it can be seen that source text is the complex sentence while target text is the simplified sentence. Sentence is simplified either by replacing a complex word with a simpler word, or by providing a definition of the complex term as in *pulmonary edema*(fluid in the lungs), *high altitude cerebral edema*. While in some cases target text is same as source text which is a bit surprising. So it is not necessary that simplified sentence is always smaller than the complex sentence but the maximum length of complex sentence is 100 while that of a simple sentence is 60 as shown in Figure 11.

Source Text	Target Text
under conditions of high humidity , the rate of evaporation of sweat from the skin decreases .	with a higher humidity , the rate of evaporation is less .
the lack of oxygen above 2,400 metres (8,000 ft) can cause serious illnesses such as altitude sickness , high altitude pulmonary edema , and high altitude cerebral edema .	this can cause illnesses such as altitude sickness , high altitude pulmonary edema (fluid in the lungs) , and high altitude cerebral edema (fluid in the brain , causing headaches and confusion) .
the human body can adapt to high altitude by breathing faster , having a higher heart rate , and adjusting its blood chemistry	the human body can deal with high altitude by breathing faster , having a higher heart rate , and changing the blood itself to have more red blood cells that can carry oxygen .
for example , hemoglobin and myoglobin contain an iron center coordinated to the nitrogen atoms of a porphyrin ring ; magnesium is the center of a chlorin ring in chlorophyll .	for example , hemoglobin and myoglobin contain an iron center coordinated to the nitrogen atoms of a porphyrin ring ; magnesium is the center of a chlorin ring in chlorophyll .
schistosomiasis , caused by one genus of trematodes , is the second-most devastating of all human diseases caused by parasites , surpassed only by malaria	schistosomiasis , caused by one genus of trematodes , is the second most devastating of all human diseases caused by parasites , surpassed only by malaria

Table 1: Sample Dataset

Readability Measure

The readability scores as discussed previously of the dataset Medical EW-SEW varied a lot. The Automated Readability Index and Flesch Kincaid score of source text varied between 10-17 while for target/simplified text was between 5-17 as shown in Table 2. But an interesting thing to note is that for some text the score of readability for target is more, some have less while some are same than the source text score. This implies that this readability measure is not reliable.

Propose New Readability Score

As shown in previous reason, previous readability measure of ARI and Flesch Kincaid Score are not reliable so the authors in

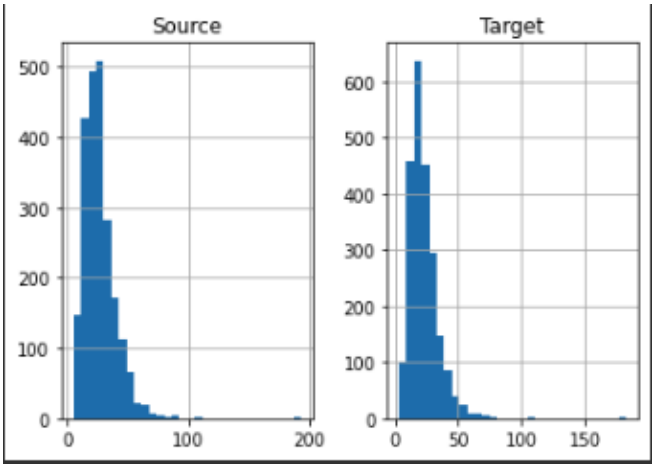


Figure 2: Dataset Lengths

Measure	Source	Target
ARI	10-17	5-17
Flesch Kincaid Score	10-17	8-17

Table 2: Readability Measure Scores

[6] proposed a new readability and technicality measure based on BERT and SciBERT Masked Language Model. It is based on the notion that as BERT [7] is trained on lay language, BERT gives more preference to simple words as compared to SciBERT [3] which is trained on scientific literature as shown in Figure 4. To measure the technicality of the source and target text, sentences were tokenized and 15% of the tokens were masked and applied masked language model to predict the masked word whose probability was stored and histogram is plotted as shown in Figure 3. The plots generated from BERT and SciBERT are illustrated in Figure 5.

METHODOLOGY

Replacement Approach

For performing the replacement technique, ordinary subset of medical EW-SEW dataset is utilized as the input document. As mentioned in Dataset section, ordinary subset consists of 2267 regular English Wikipedia sentence pairs that are filtered to contain at least one medical entity mention of type "Disease or Syndrome" and "Clinical Drug"[12]. After simplification is completed, simple subset is utilized for evaluating the results. The only difference of simple subset from ordinary subset is that simple subset consists of Simple English Wikipedia sentence pairs instead of regular English Wikipedia.

The replacement technique could be investigated in three sequential phases which are " Complex Word Identification", "Candidate Words Generation" and "Best Candidate Word Retrieval"

Complex Word Identification

To capture possible complex words in the ordinary dataset, we first extracted all the medical concepts from each sentence by utilizing MetaMap. MetaMap is a tool to discover UMLS Metathesaurus concepts referred to in given text[2].

procedure MASKED-PROB(D, M)
  sents $\leftarrow$ SENTENCE-SPLIT(D)
  $P \leftarrow$ Initialize empty list
  for $i = 1 \ldots |\text{sents}|$ **do**
    $T \leftarrow$ TOKENIZE(sents$[i]$)
    for $j = 1 \ldots 10$ **do**
      $A \leftarrow$ sample 15% from $1 \ldots |T|$
      $T' \leftarrow T$
      for all $a \in A$ **do**
        $T'[a] \leftarrow$ [MASK]
      outputs $\leftarrow$ FORWARD(M, T')
      for all $a \in A$ **do**
        prob $\leftarrow$ outputs$[a][T[a]]$
        APPEND(P, prob)
  return mean(P)

Figure 3: Technicality Measure with Masked Language Model

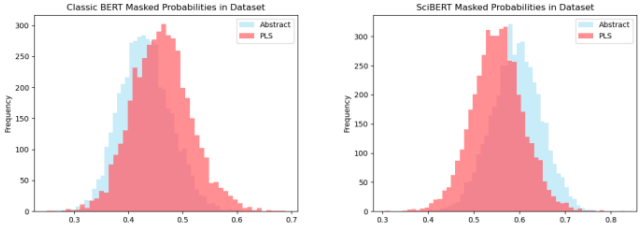
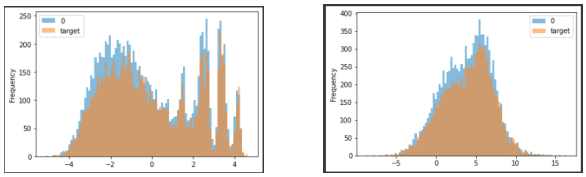


Figure 4: Expected MLM probabilities

In this paper, we used Python wrapper of MetaMap called pymetamap. To extract the medical concepts from the text, we setup pymetamap and ran it on the background. After retrieving all the medical concepts, we stored them in a separate text file.

Nevertheless, it is not a smart approach to treat all the extracted concepts as complex terms since they might include trivial words such as cause, sickness, year and so on. Hence, the resulting medical concepts are filtered in terms of their zipf frequency values. Zipf score is a logarithmic value which is measured by calculating the base 10 logarithm of the number



((a)) BERT ((b)) SciBERT

Figure 5: Masked Language Model Plots

```
# initialize metamap
mm_home = '/Users/bariskaracan/Downloads/public_mm/bin/metamap16'
mm = MetaMap.get_instance(mm_home)
```

Figure 6: A code snippet to initialize MetaMap

```
concept:  ConceptMMI(index='-e 1', mm='MMI', score='14.64', preferred_name='Myocardial Infarction'
```

Figure 7: An example of a MetaMap concept

of times the word appears per billion words (log10(wfpb)). For instance, if a word has zipf score of 1, it occurs 1 per 100 million words. According to [17], zipf scale has the tipping point from low frequency to high frequency between 3-4. Therefore, to maximize the amount of words to simplify, we also considered the words within that range as complex. At the end, we finalized identification of complex words by obtaining medical words which have zipf scores less than or equal to 4.

```
from wordfreq import zipf_frequency
zipf_frequency('heart', 'en')
```

```
5.31
```

Figure 8: Example of a simple word which has zipf score > 4

Candidate Words Generation

After identification of complex words, candidate words that are simpler than corresponding complex words are generated by utilizing BERT model's masked language modeling (MLM) property. As default, BERT is optimized by MLM task in which BERT predicts the missing tokens in a sequence given its left and right context.

For each complex word w in a sentence S, w is masked by utilizing the special token "[MASK]". Then, a vocabulary of words has been generated and their probability distribution corresponding to the masked word w is obtained. The higher probability words in the resulting vocabulary fits the context of the masked word in addition to the masked word itself[13]. Therefore, top 100 highest probability words from the vocabulary are selected as candidates for simplification. Lastly, from the chosen candidates, non-alpha numericals, stopwords and words that have higher zipf score than the masked word are removed.

In our project, MLM of pre-trained BERT-base, BioClinicalBERT and PubMedBERT models are implemented via utilizing HuggingFace libraries and compared to each other. BioClinicalBERT is trained on all notes from MIMIC 3 which is a database containing electronic health records from various patients[1]. Likewise, PubMedBERT is pre-trained from scratch using abstracts from PubMed and full-text articles from PubMedCentral[8].

Best Candidate Word Retrieval

After generating potential candidates for each BERT model, pairwise similarity scores between the list of candidate words

```
from wordfreq import zipf_frequency
zipf_frequency('pulmonary', 'en')
```

```
3.5
```

Figure 9: Example of a complex word which has zipf score < 4

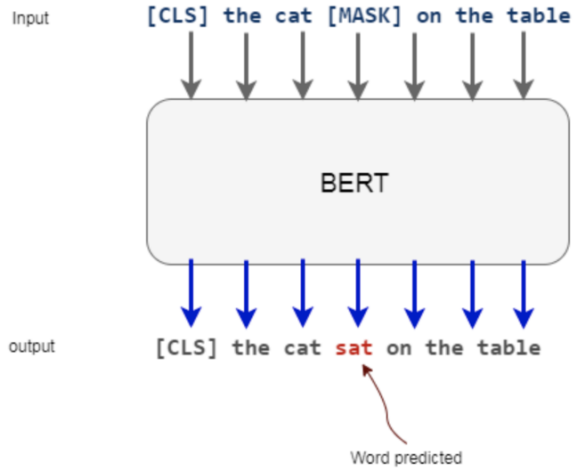


Figure 10: Representation of BERT MLM architecture for masked word prediction

and the source complex word are measured to retrieve the best candidate for replacement. For this purpose, SentenceTransformer[14] library from HuggingFace is utilized and the model "all-mpnet-base-v2" is selected to generate embeddings from the list of candidates. "all-mpnet-base-v2" model is trained on a large and diverse dataset over 1 billion training pairs and provides the best quality among HuggingFace SentenceTransformers.

Next, cosine similarity function is implemented by using sklearn library to score the pairwise similarity between the embeddings of candidates and the given complex word. Finally, the word with the highest similarity score is chosen as the best candidate for simplification and swapped with the original complex word. To sum up, we simplified the ordinary subset of Medical EW-SEW dataset by applying this technique to all of its complex words.

```
words to replace:  pulmonary          (0.81462264, 'lung')
```

Figure 11: A tuple of the best candidate word and its cosine similarity score for the complex word 'pulmonary'

Generation

Understand Style of Words

To understand and analyse style of words, we trained a classifier to classify whether the text is complex or simple giving them labels 0 and 1 respectively. The text is represented as bag of words frequency vector with vocabulary being that of BART [10] xsum dataset. 2000 training samples were used to train a Logistic Regression model which achieved an accuracy of 70% on test set of 500 samples.

Complex Word	Sample of Candidate Words
pulmonary	['cerebral', 'lung', 'respiratory', 'cardiac', 'muscular', 'systemic', 'peripheral', 'vascular', 'pedal', 'brain']
edema	['congestion', 'disease', 'swelling', 'infection', 'symptoms', 'illness', 'syndrome', 'issues', 'irritation', 'pressures']
cerebral	['brain', 'cardiac', 'peripheral', 'systematic', 'facial', 'lung', 'cognitive', 'metabolic', 'kidney', 'central']
porphyrin	['protein', 'iron', 'zinc', 'metal', 'ring', 'insulin', 'fat', 'large', 'beta', 'phosphate']
schistosomiasis	['fever', 'depression', 'hiv', 'malaria', 'aids', 'disease', 'cancer', 'plague', 'pneumonia', 'tuberculosis']

Table 3: List of 10 sample generated candidates for the given complex words (candidates are not ordered)

To better understand the tokens comprising complex sentence and simple sentence, weights of the logistic regression model was analysed and it can be clearly illustrated from Table 4 that most negative weights were assigned to complex, jargon heavy medical terms where as lay language and simple words were assigned a relatively small, positive weights. In the next section, we will use these learned logistic regression weights to penalise the generation model to not generate these complex terms and only stick to the simple terms.

Token	-ve weights	Token	+ve weights
infection	-0.822	time	0.8269
typically	-0.930	when	0.827
treatment	-0.622	some	0.898
acute	-0.8623	get	0.928
cold	-0.6753	have	0.9414
contracted	-0.5447	person	0.994
vision	-0.538	they	1.10
anterior	-0.537	it	1.002
described	-0.527	said	1.602

Table 4: Logistic Regression Weights

Generate Simplified Sentence

To generate simplified sentence from the complex sentences, two different BART models were trained. The baseline model we trained was BART which as explained in the related works also is an encoder-decoder seq2seq transformer with a bidirectional encoder and a left-right decoder. The decoder in the BART makes it appropriate for generation tasks. The baseline BART encoder-decoder model was pretrained on xsum dataset [11] (facebook/bart-large-xsum' model) provided by HuggingFace.

In the generation step, to decode- we used nucleus sampling as explained in the previous section so that at each time step, the next word generated is sampled from a probability distribution whose cumulative probability exceeds p. It ensures that words generated should be only from the head the probability mass of the output and not from the tail.

On the other hand for the second model, to encourage the model to generate simple words and discourage or penalise the model when it generates complex medical term(unlikely words), we added unlikelihood loss [18].We do this by adding a penalty term $UL = \sum_{j=1}^{|S|} -log(1 - p_\theta(s_j|y_{<t}, x))$ to the maximum likelihood loss that is usually used. The penalty term can also be weighted and then added. Here, x is the complex medical text which is given as input to the encoder, while at an instant time-step t, the output generated by the model is $y_{<t}$ while $p_\theta(s_j|y_{<t}, x)$ is the probability of token s_j and θ are the BART model parameters. The tokens s_j and their weights are obtained from training logistic regression model with BART vocabulary to classify whether the text was simple or complex as described in previous section.

EVALUATION

In this section, we will evaluate the two methods -replacement and generation used to simplify the complex medical text. We used Medical EW-SEW dataset to train the model.

Replacement

To evaluate the effectiveness of the replacement approach, first simplified sentences retrieved from each model are stored in seperate documents (text files). Next, each document is compared to simple subset of Medical EW-SEW dataset by their respective ROUGE scores. ROUGE is a set of metrics that compares automatically produced documents against a set of reference documents. In this project, we calculated three different ROUGE scores for each model which are ROUGE-1 (overlap of unigrams between the automated and reference documents), ROUGE-2 (overlap of bigrams between the automated and reference documents) and ROUGE-L (measures longest matching sequence of words).

Model	Rouge-1 Precision	Rouge-1 Recall	Rouge-1 F-measure
BERT-base	64.06	74.51	66.31
BioClinicalBERT	64.00	74.42	66.24
PubMedBERT	**64.11**	**74.58**	**66.37**

Table 5: ROUGE-1 Score Comparison of BERT models

Model	Rouge-2 Precision	Rouge-2 Recall	Rouge-2 F-measure
BERT-base	50.71	58.93	52.45
BioClinicalBERT	50.63	58.83	52.37
PubMedBERT	**50.78**	**59.04**	**52.54**

Table 6: ROUGE-2 Score Comparison of BERT models

Model	Rouge-L Precision	Rouge-L Recall	Rouge-L F-measure
BERT-base	61.65	71.53	63.80
BioClinicalBERT	61.60	71.46	63.75
PubMedBERT	**61.69**	**71.59**	**63.85**

Table 7: ROUGE-2 Score Comparison of BERT models

According to experimental results shown in Table 3,4 and 5, surprisingly all of the models performed almost identical regardless of their domain differences. Although, PubMedBERT performed better than all the models in each ROUGE score, difference is very little and negligible. Furthermore, ROUGE-2 scores appeared to be lower than ROUGE-1 and ROUGE-L.

Source Text	Predicted Text
the lack of oxygen above 2,400 metres (8,000 ft) can cause serious illnesses such as altitude sickness , high altitude **pulmonary edema** , and high altitude **cerebral edema** .	the lack of oxygen above 2,400 metres (8,000 ft) can cause serious illnesses such as altitude sickness , high altitude **lung swelling** , and high altitude **brain swelling** .
in 1920, the bucks had a daughter, carol , afflicted with **phenylketonuria** .	in 1920, the bucks had a daughter, carol , afflicted with **diabetes** .

Table 8: Model Output from BioClinicalBERT Using Replacement Approach (red: complex word, blue: simplified word)

This suggests that our replacement approach needs improvement to predict the cases where there are complex medical terms with more than one words such as 'myocardial infarction'

Finally, as shown in Table 6, our replacement method could produce both accurate and inaccurate results. In the first example, complex terms 'pulmonary edema' and 'cerebral edema' are converted into 'lung swelling' and 'brain swelling' which are simpler and reasonably accurate predictions. On the other hand, in the second example, complex term 'phenylketonuria' swapped with the simpler term 'diabetes'. Although the predicted term is simpler, these two conditions are completely different than each other. Hence, the prediction was inaccurate.

Generation

To generate simplified texts, two BART model variants were trained finetuned on xsum dataset with vocabulary size of 50,264 tokens. The Bart For Conditional Generation model was pretrained on facebook/bart-large-xsum' dataset was trained with the Medical EW-SEW dataset with source length as 100 and target lenth as 60. HuggingFace Transformers library was used to train the model keeping the batch size as 1, learning rate as 3e-5, temperature as 1.0 and top-p as 0.9 same as the initial paper [6]. Random seed has been kept 0. To discourage the model from repeating previous words, ngram value was kept 4, length penalty as 0.8. For each complex medical sentence, we generate 3 sequences. To add unlikelihood loss,a custom loss function was created with a custom seq2seqTrainer.

1000 samples were used to train the model while 1000 samples were used for validation. Colab GPU was used to train the models. To evaluate the two models, ROUGE (Recall Oriented Understudy for Gisting Evaluation) and METEOR (Metric for Evaluation of Translation with Explicit ORdering) were used which are used to evaluate Natural Language Generation(NLG) tasks such as this one. Based on different n-grams, Rouge can be of many types Rouge-1, Rouge-2, Rouge-L, RougeLSum. The different rouge and meteor scores can be seen in Table 9. As we can see, simple BART model performed much better. For the source text,the generated outputs can be seen in Table 10.

Model	Training Loss	Validation Loss	Rouge1	Rouge2	RougeL	RougeLsum	Meteor
Simple	0.4726	0.581	67.42	54.57	64.52	64.48	0.673
Unlikelihood Loss	-293.77	37.5	3.75	0	3.75	3.75	0.014

Table 9: Model Results

Source Text	Generated Outputs		
Under conditions of high humidity the rate of evaporation of sweat from the skin decreases	If the number of hours is raised in high humidity, this reduces the	During this time, the rate of evaporation of sweat from the skin lessens	If conditions of high humidity, there is higher rate of evaporation from the skin
the lack of oxygen above 2,400 metres (8,000 ft) can cause serious illnesses such as altitude sickness , high altitude pulmonary edema , and high altitude cerebral edema .	this can cause symptoms such as altitude sickness , high altitude pulmonary edema , and high altitude cerebral edema	this has some serious diseases such as altitude sickness , high altitude pulmonary edema , and high altitude cerebral edema	

Table 10: Model Output from Training BART model

FUTURE WORK

To increase the effectiveness of the proposed system, a few improvements could be done in the future for both replacement and generation approaches.

First, in replacement approach, other metrics could be used besides zipf frequency values for filtering the candidate words. Furthermore, since we utilized pre-trained BERT models to generate simplified sentences, our replacement performance is totally dependent on the performance of these models. Hence, it could be useful to train our model with the arguments we determine for obtaining better results. Besides BERT, other novel models such as T5 and GPT3 could also be utilized for our simplification system. Finally, instead of word-level masking, concept-level masking could be implemented to have better perception over complex medical terms as a whole such as "myocardial infarction" and "bubonic plague".

As we see in the dataset, that some sentences were not even simplified while in some just the definition was provided. For improving the results in generation method, we plan to come up with a better dataset that actually simplifies the complex medical terms. To map the complex terms to simpler words, we will come up with a better representation, as in logistic regressions some words lost its representation which were not present in the BART tokenizer. So, instead of logistic regression,we will come up with a better context dependent model. Secondly, we still need to figure out why unlikelihood training did not work. Thirdly, it was surprising to note how different decoding strategies generated a different meaningful output. But it was surprising to see these could not be applied while training the model. We will find a way to incorporate this in the future and also run more experiments. Lastly, we need to come up with better readability and technicality measure which is a good parameter to define simple text.

CONCLUSION

Pre-existing readability measures are inconclusive and hence incomplete to measure the readability and technicality of the text and therefore a new measure was suggested which utilizes two approaches: replacement and generation. For generation, BART model was trained on xsum dataset. It is effective in simplifying the complex sentences loaded with heavy medical jargon terms. But there is still room for improvement in terms of its effectiveness. Similarly, replacement approach achieved both successful and unsuccessful results while swapping complex words with simpler ones. Because of utilizing pre-existing language models, its performance is mostly dependent on the performance of these models for candidate generation.

REFERENCES

[1] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323* (2019).

[2] Lan Aronson. 2020. MetaMap - A Tool For Recognizing UMLS Concepts in Text. (2020).

[3] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3615–3620.

[4] Gretchen K Berland, Marc N Elliott, Leo S Morales, Jeffrey I Algazy, Richard L Kravitz, Michael S Broder, David E Kanouse, Jorge A Muñoz, Juan-Antonio Puyol, Marielena Lara, and others. 2001. Health information on the Internet: accessibility, quality, and readability in English and Spanish. *jama* 285, 20 (2001), 2612–2621.

[5] Asad J Choudhry, Yaser MK Baghdadi, Amy E Wagie, Elizabeth B Habermann, Stephanie F Heller, Donald H Jenkins, Daniel C Cullinane, and Martin D Zielinski. 2016. Readability of discharge summaries: with what level of information are we dismissing our patients? *The American Journal of Surgery* 211, 3 (2016), 631–636.

[6] Ashwin Devaraj, Iain Marshall, Byron Wallace, and Junyi Jessy Li. 2021. Paragraph-level Simplification of Medical Texts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 4972–4984.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186.

[8] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)* 3, 1 (2021), 1–23.

[9] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The Curious Case of Neural Text

Degeneration. In *International Conference on Learning Representations*.

[10] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 7871–7880.

[11] Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 1797–1807.

[12] Nikhil Pattisapu, Nishant Prabhu, Smriti Bhati, and Vasudeva Varma. 2020. Leveraging Social Media for Medical Text Simplification. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 851–860.

[13] Jipeng Qiang, Yun Li, Yi Zhu, Yunhao Yuan, Yang Shi, and Xindong Wu. 2021. LSBert: Lexical Simplification Based on BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 3064–3076.

[14] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. `https://arxiv.org/abs/1908.10084`

[15] Tarek Sakakini, Jong Yoon Lee, Aditya Duri, Renato FL Azevedo, Victor Sadauskas, Kuangxiao Gu, Suma Bhat, Dan Morrow, James Graumlich, Saqib Walayat, and others. 2020. Context-Aware Automatic Text Simplification of Health Materials in Low-Resource Domains. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*. 115–126.

[16] E A Smith and R. Senter. 1967. Automated readability index. *AMRL-TR. Aerospace Medical Research Laboratories* (1967), 1–14.

[17] Walter JB Van Heuven, Pawel Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2014. SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly journal of experimental psychology* 67, 6 (2014), 1176–1190.

[18] Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2019. Neural Text Generation with Unlikelihood Training. (2019).