Get to the Point! A Simplification-Based Medical Question-Answering Bot

Hemanth Venkata Kota
hvkota

Adithya Shankar Bhattiprolu
abhattip

Abstract

Interpreting complex documents filled with technical jargon can be challenging for individuals who lack deep insight into a particular field. This difficulty is especially prominent in domains such as medicine and law. To address this issue, our project aims to build upon the English text simplification work presented in [1] by incorporating two new medical datasets for evaluation. In our approach, we employ state-of-the-art models such as Flan-T5, GPT-2, and Pegasus to tackle the task of medical text simplification. By leveraging these advanced models and conducting a comprehensive analysis, our project aims to enhance the field of text simplification, particularly in the medical domain.

1 Introduction

Human health relies on intricate biological systems that are often described using complex and hard-to-remember scientific terms. However, it is crucial for the well-being of individuals that even the general public can comprehend and remember health reports. This necessity has inspired us to undertake the task of simplifying medical technical documents. By making these documents more accessible and easier to understand, our ultimate aim is to bridge the gap in accessibility and trust between the medical field and under-resourced individuals. We believe that everyone should have the opportunity to access and comprehend vital medical information, regardless of their background or available resources.

There are two primary approaches to text simplification: lexical simplification, which involves paraphrasing or replacing complex words, and structural simplification, which focuses on removing difficult words altogether. However, the main challenge in supervised learning for text simplification lies in the limited availability of training data. Efforts like Med-Easi [2] are actively addressing this issue by addressing the shortage of training data.

In a recent study by [3], Cochrane was utilized as a resource to train different models such as BERT and ControlTS. In our work, we aim to conduct a comprehensive analysis of various language models for text simplification across multiple datasets, evaluating their performance using diverse evaluation metrics. By examining these models in detail, we strive to advance the field of text simplification and improve its effectiveness.

We strive to make complex medical information more accessible and understandable for individuals who may not possess extensive domain-specific knowledge. Initially, we follow the guidelines outlined in the paper to set up a zero-shot learning framework. Subsequently, we extend this framework to analyze the performance of the models on Med-Easi and Cochrane datasets, which serve as valuable benchmarks for evaluation.

This paper is structured as follows: Section 2 provides an overview of the existing work in the field of text simplification and summarization, highlighting the relevant research and approaches. In Section 3, we delve into our approach to addressing the problem at hand. We outline the methodologies and techniques employed in our study, explaining the rationale behind our choices. Section 4 focuses on the experiments conducted and the results obtained. We provide a detailed account of the experimental setup, including the datasets used, evaluation metrics employed, and the outcomes of our analysis. Finally, in Section 5, we present our future directions for research and draw conclusions based on our findings. We discuss potential avenues for further exploration and development in the field of text simplification and summarization.

2 Related work

Our work began by exploring medical datasets created by researchers such as [2] and [3]. These studies share a common focus on addressing the challenge of creating large medical datasets specifically designed for text simplification. In the supervised learning setting, reinforcement learning models [4] are employed to generate sentence key pairs for constructing medical datasets. Previous studies have utilized various models, including neural machine translation [5] and pretrained BART [3], for supervised training.

In parallel, the research field also explores unsupervised neural text simplification [6] to tackle the issue of dataset scale. This approach involves a shared encoder and a pair of attention-based decoders. To evaluate the performance of our models, we applied fine-tuned language models such as Flan T5, Pegasus, and GPT-2, which outperformed the aforementioned models based on metrics like SARI and BLUE scores.

Our contributions primarily lie within the supervised setting, expanding upon the existing work of researchers like [7], [**teslea**]. Our approach involves taking a complex medical brief and utilizing zero-shot learning to generate simplified text while adhering to given constraints. By incorporating these advancements, we aim to enhance the field of text simplification in the medical domain and contribute to the broader body of literature on simplification techniques.

3 Approach

In this section we present our approach for assessing various zero-shot and fine-tuned models across 2 different datasets using 5 distinct evaluation metrics.

3.1 Datasets

Acquiring medical data is a challenging task due to the strict privacy regulations imposed by HIPAA. However, for our approach, we were able to utilize two datasets: MedEasi **[2]** and the Medical Cochrane dataset.

The MedEasi dataset comprises 3600 pairs of (text, simplified text), providing valuable resources for text simplification tasks. On the other hand, the Medical Cochrane dataset consists of 4259 such pairs and was obtained by scraping the pubwiki website. This dataset is curated by NHS healthcare professionals and experts, aiming to deliver high-quality, independent, and evidence-based healthcare information. It covers a wide range of health issues, including mental health, gynaecology, oncology, and more. Notably, it offers detailed information on specific topics, such as oncology, which includes 426 use cases as mentioned in the article **[8]**. These datasets serve as valuable resources for our research, enabling us to train and evaluate our models on real-world medical text, and encompassing a broad spectrum of healthcare topics and expert-driven information.

3.2 Models

We used the following models for the text simplification task

1. **FLAN-T5 (Zero Shot):** Flan-T5 is a text-to-text transformer model based on an encoder-decoder architecture, designed for performing natural language processing tasks. In our work, we utilized the flan-t5-base model provided by the Hugging Face library, which serves as a robust and reliable resource for our text processing needs.[9]

2. **FLAN-T5 (Fine Tuned):** In this setup, we finetune the Flan-t5-base model using the medical Cochrane training dataset (1243 samples) and evaluate its performance on test set.

3. **Pegasus (Zero Shot):** The PEGASUS model is a state-of-the-art sequence-to-sequence model for abstractive text summarization. In our work, we leveraged the pegasus-xsum model provided by the Hugging Face library. This particular model, optimized for summarization tasks, served as a powerful tool for our text simplification goals.

4. **Pegasus (Fine Tuned):** Similar to Flan-T5 we finetune the model using the training set of medical cochrane.

5. **GPT-2 (Zero Shot):** GPT-2 is a transformer with 1.5 billion parameters trained over 8 million websites. It has 10 hyperparameters to fine tune. We use the hugging face gpt-2 model for comparitive studies.

6. **Keep it Simple (KiS) (Zero Shot):** KiS [10] is an Unsupervised text simplification model which optimizes over three properties: fluency, salience and simplicity of a language.

We used the "keep-it-simple" model that is present in huggingface. The model was pretrained on news dataset.

For generating sentences, we used the following format -

$$"Simplify : Medical\ Extract"$$

We conducted experiments to fine-tune both GPT-2 and KIS models. However, during the fine-tuning process, we encountered challenges with GPT-2. The model failed to generate coherent sentences as desired, resulting in outputs that did not meet our evaluation criteria. As a result, we decided not to include GPT-2 in our final evaluation. Similarly, with the KIS model, we faced difficulties in resolving certain dependency issues, specifically related to deprecated packages. These unresolved issues prevented us from successfully training the KIS model for our specific task. Despite our best efforts, we were unable to overcome these challenges within the scope of our project. Given these limitations, we proceeded with alternative models that provided satisfactory performance and reliability for our evaluation.

3.3 Metrics

For the evaluating the performance of the models that were used we used the following evaluation metrics -

1. **BLEU -** BLEU is a metric used for comparing a generated sentence to a reference sentence. A higher BLEU score indicates better similarity between the generated and reference texts.

2. **SARI -** SARI is a metric used to evaluate the quality of text paraphrasing or rewriting systems by comparing the generated text to the reference text. A higher SARI score indicates better quality and similarity to the reference paraphrases. [keep,add,delete] are the operation over which it is calculated.

3. **ROUGE -** We consider two varitions of ROUGE,Rouge L (longest common subsequence) and Rouge 1 (unigram level overlap) for our evaluation.

4. **BertScore -** BERTScore computes a similarity score based on the cosine similarity between the BERT embeddings of words or subwords in the generated and reference texts.

We initially planned on using the Flesch-Kincaid Reading Ease Score [11] but then decided not to because of the ease with which it can be manipulated as described in [12].

4 Experiments

4.1 Cochrane Simplification

In this subsection, we evaluate the performance of all the models described in section 3 on the medical cochrane (M-cochrane) dataset. Table 1 shows the results obtained. In the table, the entry "Baseline" denotes the value of the metrics when evaluated on the test set of medical cochrane (using the original sentence and the simplification provided in the dataset).

Based on the table, it is evident that the fine-tuned models (Flan-T5 and Pegasus) as well as the GPT-2 (ZS) and (KiS) models outperformed the baseline. Notably, the KiS model and GPT-2 exhibited substantial improvements over the baseline. However, upon closer examination of the generated sentences produced by these models, we discovered an issue with the evaluation metrics. The GPT-2 model, while successful in simplifying the text to a certain extent, suffered from a phenomenon known as "hallucination" **https://rdcu.be/decdb**. It began adding random facts and names to the generated text that were not present in the original prompt. This compromised the accuracy and coherence of the generated content. On the other hand, the KiS model failed to effectively simplify the text. Instead, it memorized the original prompt without introducing any simplifications, resulting in output that was identical to the input prompt. As a consequence of these limitations, we made the decision to exclude these models from further experiments since we were unable to successfully fine-tune them. Table 4 illustrates an instance of generation for each of the models, further highlighting these observations.

We additionally computed the BertScore for the finetuned models to see if the produced simplified results is relevant to the prompt that was generated. From table 2 we can see that the fine-tuned FLAN-T5 model is able to generate the most

Model	BLEU	R1	RL	SARI
FLAN-T5 (FT)	13.43	44.97	43.77	36.57
Pegasus (FT)	10.75	43.34	28.42	39.51
Baseline	8.70	43.12	24.79	-
FLAN-T5 (ZS)	3e-6	7.7	6.9	31.55
Pegasus (ZS)	7e-4	12.7	10.2	33.60
GPT-2 (ZS)	73.22	82.96	82.96	41.52
KiS (ZS)	**99.45**	**98.73**	**98.73**	**82.46**

Table 1: Performance on M-Cochrane

Model	BertScore
FLAN-T5 (FT)	**90.28**
Pegasus (FT)	86.18
Baseline	85.75

Table 2: Bertscore for M-Cochrane

relevant simplifications when compared to pegasus and the baseline. The BertScore for GPT-2 and KiS was also computed and a similar issue as above was observed, where the memorization caused an erroneous BertScore.

4.2 MedEasi Simplification

We conducted a cross-domain evaluation of the finetuned models to assess their performance across different domains. The objective of this experiment was to determine whether the models could maintain their zero-shot capabilities and demonstrate good generalization on unseen data. To evaluate the models, we tested the previously finetuned models trained on the medical Cochrane dataset on the relatively simpler medEasi dataset. The results obtained are presented in Table 3. Comparing the performance of the finetuned models to their zero-shot counterparts, we observed improvements in their performance. However, it is worth noting that the pegasus model was unable to surpass the baseline in this particular scenario. It is possible that additional steps or adjustments are necessary to further enhance the model's performance in this domain.

4.3 Miscellaneous Experiments

The following three experiments were conducted to evaluate the performance of the flan-T5 model.

4.3.1 Effect of Training Epochs

We conducted experiments to explore the impact of varying the number of training epochs during the fine-tuning process of Pegasus and Flan-T5

Model	BLEU	R1	RL	SARI
FLAN-T5 (FT)	**71.44**	**87.19**	**86.04**	**42.76**
Pegasus (FT)	9	24.37	21.68	27.96
Baseline	36.43	63.67	60.87	-
FLAN-T5 (ZS)	68.83	85.51	83.59	42.54
Pegasus (ZS)	7.8	28.67	24.20	28.28

Table 3: Performance on Medeasi

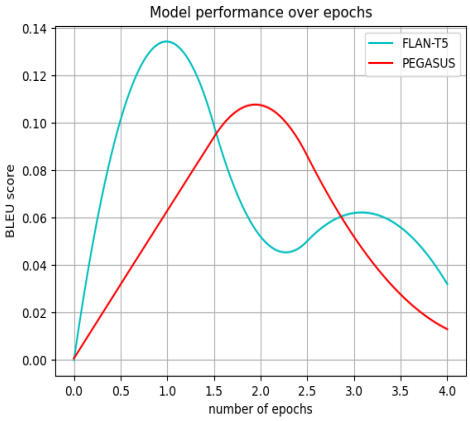

Figure 1: Variation in BLEU score as the number of epochs vary for Flan-T5

models. The objective was to understand how different epoch settings affect the evaluation metrics. Figure 1 illustrates the relationship between the number of training epochs and the model's performance. Our findings revealed an interesting trend: initially, as the number of epochs increased, the model's performance improved. However, beyond a certain point, we observed a decline in performance. Upon inspecting the generated sentences, we noticed a significant degradation in quality, characterized by a high number of repeating tokens. Furthermore, after 3 epochs, the generated sentences lost their coherence. Hence, for all the experiments we train the model as long as the performance improves and stop training if it degrades. (1 epoch for Flan-T5 and 2 epochs for Pegasus model)

4.3.2 Variation of Beam Size

We employed the beam search algorithm as the decoding mechanism for sentence generation in both Flan-T5 and Pegasus models. In this particular experiment, we selected the best performing model (Flan-T5 trained over 1 epoch) as a reference point. Our objective was to investigate the impact of varying the number of beams used dur-

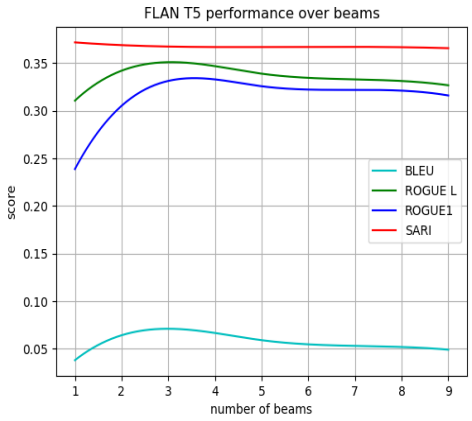

Figure 2: Variation in Different Metrics as the number of beams vary for Flan-T5

ing the decoding process and observe its effect on the evaluation metrics.

Figure 2 denotes the change in all the evaluation metrics for the flan-T5 model when the number of beams is varied. We see that as the number of beams increases most of the metrics saturate at a particular value. Better performance was observed when the beam size was set to 2 (considering time taken for training the models as well) hence the value is set to 2 for all experiments.

Another possible reason for why metrics like BLEU do not increase is that As the number of beams increases, the search space expands, allowing for more diverse candidate sequences. However, this increased diversity can also introduce more noise and lead to a decrease in the quality of the generated sentences. Each beam represents a hypothesis, and during the search process, the model selects the most probable continuation at each step based on the scoring criteria. With a larger number of beams, the model explores a wider range of possibilities, potentially including sub-optimal or incorrect choices. This can result in lower BLEU scores, as BLEU measures the similarity between the generated sentences and the reference sentences.

4.3.3 Remaining Model Parameters

The remaining generation parameters such as temperature, repeat penalty and length penalty were also varied to observe its impact but no noticeable difference was observed in the fine-tuned models. The learning rate was determined using a manual grid search to obtain a suitable value for the remaining experiments (5 x e^{-4} for the finetuning).

4.3.4 Dataset Bias

In this subsection, we present a data analysis of the obtained BLEU scores for the finetuned models. Figure 3 displays the frequency distribution of the BLEU scores specifically for the FLAN-T5 model. Similar distributions were observed for both the original distribution and the Pegasus fine-tuned dataset.

The distribution exhibits an exponential decay pattern, with the majority of the samples yielding low BLEU scores, typically falling within the range of 0-5%. These samples contribute to the overall drop in the average BLEU score.

Upon further investigation, we examined all the samples within this particular range and made an interesting observation. In approximately 85% of the cases, we found that the difference in sentence length (measured by the number of tokens) exceeded 55 words. It was noted that in all these cases, the generation process by the model was prematurely terminated, with the end-of-sentence token generated in the middle of a sentence. We hypothesize that this significant difference in sentence length is a major contributing factor to the reduced BLEU score.

For future work, we propose potential solutions to address this issue. One approach could involve preventing early stopping during sentence generation, allowing the model to produce complete and coherent sentences. Additionally, if the generated sentence is of a significantly small length, an alternative strategy could involve regenerating a sentence for the same prompt rather than moving on to the next prompt. Implementing these measures has the potential to mitigate the impact of sentence length discrepancies on the BLEU scores and improve overall performance.

5 Additional Evaluation Attempts

To address the hallucination issue encountered by GPT-2 during sentence generation for medical Cochrane, we conducted a survey aimed at gathering insights and feedback. The survey involved presenting participants with four randomly selected prompts from the medical Cochrane dataset. For each prompt, we provided four possible simplifications: the original supervised simplification, simplification generated by ChatGPT, simplification generated by Flan-T5, and simplification generated by Pegasus. Unfortunately, because of lack of time we were unable

Index	Model	Simplification	BLEU
1	FLAN-T5 (FT)	We included eight trials in this review with a total of 1708 participants. Trials were conducted in India, Nepal and South Africa. Follow-up ranged from one day to six months, but most trials reported at six to eight weeks after surgery. No studies reported presenting visual acuity so data were collected on both best-corrected (BCVA) and uncorrected (UCVA) visual acuity. Seven studies (1223 participants) reported visual acuity of 6/18 or better (rather than 6/12 or better) so this was used as an indicator of good functional vision. Seven studies (1223 participants) reported visual acuity of 6/18 or better at six"	0.04
2	PEGASUS (FT)	and follow- up to three months after surgery included eight trials with a total of 1708 participants were included in this review. randomised controlled surgical procedures. the evidence is current to six months post-to-three months after MSICS, but most trials did not report on long-term follow-up. The studies were conducted in India, Nepal, and South Africa. We found that phacoemulsification may result in better visual acuity (BCVA) in the short term (up to three months after surgery) compared to MSICS. There was no difference between the two techniques in terms of complication rates or cost. In view of the lower cost of MSICS, this may be a favourable technique in the patient populations examined in	0.017
3	GPT-2 (ZS)	Acknowledgments We thank Dr. S. S. K. Singh for his assistance in the design of the study and Dr. S. S. K. Singh for his assistance in the design of the study. We also thank Dr. S. S.	0.83
4	KIS (ZS)	**Memorize:** Long Original Prompt	0.999

Table 4: Variation of Sentences generated by Decoding algorithms.

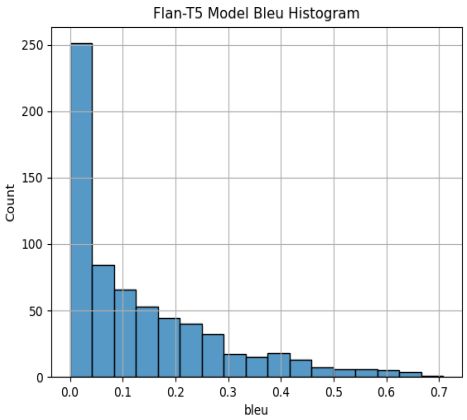

Figure 3: BLEU score Prompt Distribution

to get sufficient responses for statistical significance. A copy of the form can be found here - **https://forms.gle/Khcpc19WQT2reMCaA**.

6 Conclusion and Future work

In conclusion, our project aims to address the challenge of interpreting complex medical documents by leveraging text simplification techniques. We build upon existing work in English text simplification by incorporating two new medical datasets for evaluation. Our experiments on the MedEasi and Medical Cochrane datasets demonstrate the effectiveness of our approach. The fine-tuned models, Flan-T5 and Pegasus, outperformed the baseline in terms of various evaluation metrics such as BLEU, SARI, ROUGE, and BertScore. However, we encountered limitations with the GPT-2 and KiS models, including hallucination and lack of simplification, respectively. Additionally, during the middle of the quarter, we considered using Flesch-Kincaid as an evaluation metric for our models. However, after conducting further research, we decided to exclude it from our analysis because of the reasons cited in section 3. We had originally planned to utilize GPT models for the medical dataset; however, due to computational constraints, we were unable to train them.

However, we encountered certain limitations with the GPT-2 and KiS models. The GPT-2 model sometimes produced hallucinatory responses, while the KiS model lacked simplification capabilities. In the middle of the quarter, we initially considered using Flesch-Kincaid as an evaluation metric for our models. However, after conducting further research, we decided to exclude it from our analysis due to the reasons explained in section 3. Originally, we had planned to utilize GPT models for the medical dataset, but unfortunately, we faced computational constraints that prevented us from training them.

In the future, we plan to explore additional datasets, refine our models, and investigate novel approaches to further enhance text simplification in the medical domain. Ultimately, we strive to make a positive impact on the accessibility and comprehensibility of medical information for all individuals. You can find our implementation and other supplementary material like datasets, results, survey-form etc at Github.

References

[1] S. Joseph, K. Kazanas, K. Reina, V. J. Ramanathan, W. Xu, B. C. Wallace, and J. J. Li. Multilingual simplification of med-

ical texts, 2023. arXiv: 2305.12532 [cs.CL].

[2] C. Basu, R. Vasu, M. Yasunaga, and Q. Yang. Med-easi: finely annotated dataset and models for controllable simplification of medical texts, 2023. arXiv: 2302.09155 [cs.CL].

[3] A. Devaraj, I. J. Marshall, B. C. Wallace, and J. J. Li. Paragraph-level simplification of medical texts, 2021. arXiv: 2104.05767 [cs.CL].

[4] L. van den Bercken, R.-J. Sips, and C. Lofi. Evaluating neural text simplification in the medical domain. In *The World Wide Web Conference*, WWW '19, pages 3286–3292, San Francisco, CA, USA. Association for Computing Machinery, 2019. ISBN: 9781450366748. DOI: 10.1145/3308558.3313630. URL: https://doi.org/10.1145/3308558.3313630.

[5] T. Wang, P. Chen, J. Rochford, and J. Qiang. Text simplification using neural machine translation. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 4270–7271, Phoenix, Arizona. AAAI Press, 2016.

[6] S. Surya, A. Mishra, A. Laha, P. Jain, and K. Sankaranarayanan. Unsupervised neural text simplification, 2019. arXiv: 1810.07931 [cs.CL].

[7] R. Sun, H. Jin, and X. Wan. Document-level text simplification: dataset, criteria and baseline, 2021. arXiv: 2110.05071 [cs.CL].

[8] M. Goldkuhle, V. Narayan, A. Weigl, P. Dahm, and N. Skoetz. A systematic assessment of cochrane reviews and systematic reviews published in high-impact medical journals related to cancer. *BMJ Open*, 8:e020869, Mar. 2018. DOI: 10.1136/bmjopen-2017-020869.

[9] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, A. Castro-Ros, M. Pellat, K. Robinson, D. Valter, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, and J. Wei. Scaling instruction-finetuned language models, 2022. arXiv: 2210.11416 [cs.LG].

[10] P. Laban, T. Schnabel, P. Bennett, and M. A. Hearst. Keep it simple: unsupervised simplification of multi-paragraph text, 2021. arXiv: 2107.03444 [cs.CL].

[11] M. Solnyshkina, R. Zamaletdinov, L. Gorodetskaya, and A. Gabitov. Evaluating text complexity and flesch-kincaid grade level. *Journal of Social Studies Education Research*, 8:238–248, Nov. 2017.

[12] T. Tanprasert and D. Kauchak. Flesch-kincaid is not a text simplification evaluation metric. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 1–14, Online. Association for Computational Linguistics, Aug. 2021. DOI: 10.18653/v1/2021.gem-1.1. URL: https://aclanthology.org/2021.gem-1.1.