MedTranslate: Improving Patient-Physician Communication in Third-World Hospitals

Yoonsoo Nam, Jadrian Tan, Alfred Chen, Wilson Tan, Scott Susanto
Thomas Lord Department of Computer Science
University of Southern California

1 Abstract

Medical text simplification is critical for enhancing health literacy, particularly in under-resourced regions (Hendawi et al., 2022). This study builds upon the MedEasi corpus and the ctrlSIM model, utilizing a T5-Large model for the elaborative and abstractive simplification of medical texts (Basu et al., 2023). Recognizing the computational limitations in third-world healthcare settings, we implemented a knowledge distillation approach using the T5-Small model as a student to emulate the teacher model's proficiency. Our methodology involved a fine-tuned T5-Large with the MedEasi dataset, followed by distillation to the T5-Small model. The performance of the student model was evaluated using metrics such as SARI and ROUGE scores, along with readability tests like Flesch Readability Ease and Flesch-Kincaid Grade Level. Despite satisfactory results on these conventional metrics, human evaluations highlighted a gap in simplification effectiveness, with the student model often reproducing complex medical jargon (Phatak et al., 2022). Our research reveals the inadequacy of current metrics to fully capture the nuances of medical text simplification and underscores the need for more user-centered evaluation methods. We discuss the implications of these findings for computational efficiency in text simplification and propose future directions for developing more effective evaluation frameworks and model configurations. The data and code are available at `https://github.com/yoonsoo1/MedTranslate`.

2 Introduction

Health literacy is paramount in ensuring that patients can make informed decisions about their care (Kauchak and Leroy, 2016). Yet, the complexity of medical language often poses a barrier, particularly in resource-constrained environments where simplification could be most beneficial. This challenge is amplified in third-world settings, where there is a pressing need for efficient tools that can operate on limited computational resources.

The MedEasi paper has made significant strides in this domain by introducing a dataset that focuses on the Elaborative and Abstractive Simplification of medical texts. This dataset provides a corpus compiled from expert and simple text pairs sourced from SIMPWIKI and MSD, enriched through an annotation process that leverages the expertise of medical professionals and layperson understanding, with an additional layer of web-crawled data for comprehensiveness (Basu et al., 2023).

In our study, we adopt the T5-Large model from MedEasi as our teacher model, and through a collective decision within our research group, we implement a T5-Small model as the student model in our knowledge distillation process. This decision was motivated by the desire to develop a tool that is not only effective but also efficient enough to be deployed in settings with limited computational resources, such as small clinics in developing countries.

Our methodology involved utilizing the T5-Large teacher model with the MedEasi dataset and subsequently distilling its capabilities into the T5-Small student model. We undertook a comprehensive evaluation of the student model's performance against established metrics such as SARI and ROUGE scores, and readability tests including Flesch Readability Ease (FRE) and Flesch-Kincaid Grade Level (FKGL) (Leroy et al., 2013). Furthermore, these quantitative assessments were contrasted with human evaluations to determine the practical efficacy of the text simplifications.

Our findings point towards a satisfactory performance of the student model on some conventional metrics. However, they also reveal a discrepancy when it comes to human evaluations, suggesting a divergence between metric-based evaluations and user-centric outcomes. This critical insight under-

scores the need for continued innovation in the evaluation of medical text simplification, a theme that we discuss in depth in our conclusion and future work sections.

As we delve into the intricacies of medical text simplification, our research offers a window into the potential of smaller models in knowledge distillation and the continuous search for the optimal balance between computational efficiency and the quality of simplification.

3 Method

Knowledge distillation (Wang et al., 2020) is a machine learning technique where knowledge is transferred from a large, complex model teacher model to a smaller, simpler student model. Instead of directly learning from the training data, student model attempts to "mimic" teacher model's outputs, in forms of logits or probabilities, to optimize performance. Consequently, student model is more efficient to run due to a simpler architectural structure and a lower computational power and memory requirement.

4 Experiments

4.1 Dataset

The MedEasi dataset underpins our study, selected for its elaborative and abstractive simplification of medical texts. We utilized the SIMPWIKI and MSD corpora, chosen for their varied Levenstein Similarity and Compression ratios, ensuring a broad spectrum of text complexities (Basu et al., 2023). This rigorously curated dataset comprises 1979 expert-simple text pairs, spanning diverse medical subdomains. Its annotations facilitate our knowledge distillation from the T5-Large teacher model to the T5-Small student model, providing a rich ground for evaluating simplification efficacy against human readability standards.

4.2 Teacher & Student Model

Using the pretrained T5-Large model (Raffel et al., 2020), we started by fine-tuning the teacher model with the medical corpus from the MedEasi paper (Basu et al., 2023). We follow the fine-tuning process under the control-free simplification specified in Basu et al. (2023). Facing limited compute, the controllable simplification model with a slight higher SARI (40.89 for controllable compared to 39.28 for control-free) score was not deemed worth testing for our testing of light-weight knowledge distillation. Using the fine-tuned T5-Large model with 770 million parameters as the teacher model, we fine-tune the student model of T5-Small of only 60 million parameters (Raffel et al., 2020).

4.3 Knowledge Distillation Objective

Successful knowledge distillation for the newly introduced Transformer architecture has been difficult. Traditionally attempted logit wise distillation by minimizing the difference in prediction between the teacher and student model (Hinton et al., 2015) has been deemed inefficient in the recent architecture involving attention matrices (Wang et al., 2020).

For our method, we use a modified version of the knowledge distillation objective Wang et al. (2020) suggested for knowledge distillation of Transformer based models. (Wang et al., 2020) suggested two distillation objectives: minimize the difference between the last layer of attention heads for teacher and student models and minimize the difference between the value relations transfer. Because we use a model that has the Transformer architecture of Encoder to Decoder model, we need a way to not only distill attention heads of the last layer of encoder but also that of the decoder.

To achieve this, our proposed objective is as follows:

$$\mathcal{L}_{ENC} = \frac{1}{A_h^S|x|}\Sigma_{i=1}^{A_h^S}\Sigma_{t=1}^{|x|}D_{KL}(A_{E,a,t}^T||A_{E,a,t}^S)$$

$$\mathcal{L}_{DEC} = \frac{1}{A_h^S|x|}\Sigma_{i=1}^{A_h^S}\Sigma_{t=1}^{|x|}D_{KL}(A_{D,a,t}^T||A_{D,a,t}^S)$$

$$\mathcal{L}_{total} = \mathcal{L}_{ENC} + \mathcal{L}_{DEC}$$

where A_h^S is the number of attention heads of the student, $|x|$ is the length of the sequence, D_{KL} is Kullback–Leibler divergence (Kullback and Leibler, 1951), A^T is the teacher model's attention distribution, A^S is that of the student distribution. D and E stands for Decoder and Encoder, respectively. We calculate the attention distribution using the softmax of attention matrices:

$$A_i^Z = \frac{e^{Attention(Q,K,V)_i}}{\Sigma_j e^{Attention(Q,K,V)_j}}$$

We forgo the value relations transfer in favor of decoder attention head difference minimization.

As the information of value relations would have been encoded in the last layer of the decoder, usage of value relations for distillation would cause an overfitting to the value matrix of the input to the model. Finally, we minimize the $\mathcal{L}_{total}$ to distill the teacher model to the student model.

5 Results & Discussion

We used several evaluation metrics to measure our models performance: SARI scores, ROUGE scores, Flesch Readability Ease (FRE), Flesch-Kincaid Grade Level (FKGL), and human evaluation. ROUGE and SARI are two common metrics used in text simplification, while FRE and FKGL are two metrics that measure readability using average sentence length and average syllables per word (Kincaid et al., 1975).

We were able to obtain SARI scores (See table 1) similar to other models, but unable to get comparable FKGL scores (See table 3) on the validation set (Basu et al., 2023). We include FRE for a broader assessment.

	SARI	
Model	Validation	Test
Teacher	42.909	45.762
Student	40.962	41.450

Table 1: SARI scores for the teacher and student models.

5.1 Automatic Evaluation

The student model performs exceptionally well on ROUGE evaluation and even better than the teacher model (See table 2) which shows strong similarity between the original medical texts to the student simplified texts, but these values are suspiciously high. The high ROUGE scores could suggest that the student simplified sentences and the simplified medical texts are, word for word, especially similar which means that the student model is only performing minimal simplification. Another possibility that explains the high ROUGE values could be that the model is simplifying medical terminology and reiterating the easily understood terminology. To test this explanation, we evaluate using FRE scores. From the low FRE scores (See table 3), we suspect that the student does not appropriately simplify the original texts and only repeats them.

To explore the student model's performance on ROUGE and FRE evaluations, we also evaluate the teacher model. If the teacher model shows similar results, it would better explain the student results. Although the teacher ROUGE values show a moderate overlap in words between the teacher simplified texts and the dataset's simplified texts (See table 2), it produces texts that are just as understandable as the dataset's simplified texts based on their similar FRE scores (See table 3).

5.2 Human Evaluation

From the metrics, the student model does not effectively simplify medical text, but it is unclear what the issue is. From human evaluation, it confirms our suspicion that the student will commonly reiterate the original text, but it also successfully simplifies in certain cases. It also shows potential reasons for why it has difficulties. It is worth noting that the student performs better than the teacher on some sentences.

One possible explanation is the student model fails to fully understand the prompt. The reiteration behavior of the student model is similar to that of the unprompted teacher model, except the student will occasionally attempt to translate to German and usually outputs the "$expert$" prompt which the fine-tuned teacher model excludes. Overall, the student tries to mimic behavior of the unprompted teacher, the prompted teacher, and the base T5-Small. If the model is failing to fully learn the prompts, it could explain why the model mimics the unprompted teacher model and only performs minimal simplification. This also explains why the student model will attempt to translate some words to German (See table 4) in several datapoints which is what the base T5-Small model does. This may be due to the type of knowledge distillation used or simply because the model is too small to capture the prompt's meaning. Fine-tuning the student model to learn the prompts could also be beneficial.

6 Conclusion

6.1 Current State

Using currently available metrics, our student model appears to be performing to a satisfactory standard on some metrics, scoring well on all of SARI, ROUGE, but not on FRE, and FKGL. Optimizing on these metrics, we can conclude that knowledge distillation was a somewhat satisfactory method for reducing larger models into smaller models that would be easier to run on smaller GPUs in third world hospitals.

On the other hand, human evaluation shows that

	ROUGE-1		ROUGE-L	
Texts	Validation	Test	Validation	Test
Dataset	0.544	0.527	0.499	0.483
Teacher	0.685	0.701	0.660	0.677
Student	0.807	0.801	0.823	0.818

Table 2: ROUGE f1-scores when comparing the dataset's unsimplified medical texts to the dataset's simplified texts, teacher simplified texts, and student simplified texts.

	Flesch Readability Ease		Flesch-Kincaid Grade Level	
Texts	Validation	Test	Validation	Test
Unsimp Dataset	36.826	39.646	13.028	12.258
Simp Dataset	48.106	53.101	11.572	10.239
Teacher	48.936	50.792	10.715	10.432
Student	31.318	39.282	13.212	11.764

Table 3: FRE and FKGL scores for the dataset's unsimplified (Unsimp Dataset) and simplified (Simp Dataset) medical texts, teacher simplified texts, and student simplified texts.

Example	Expert	Student Simple
Translation	With treatment, less than 0.2 % of children require endotracheal intubation.	Bei der Behandlung erfordert weniger als 0,2 % der Kinder eine endotracheal intubation.
Repetition	Self-harm is most common between the age of 12 and 24.	Self-harm is most common between the age of 12 and 24.
Simplified	Cystic lung disease and recurrent spontaneous pneumothorax may occur. These disorders can cause pain and shortness of breath.	Cystic lung disease and recurrent spontaneous pneumothorax may occur.

Table 4: Examples of cases when the student model translated to German, repeated the expert (unsimplified medical text), and successfully simplified. Prompts were removed in these examples.

both student and teacher model does not do a satisfactory job of decomplicating complex medical jargon into layman terms understandable by third world patients. We then learned that the current methods of evaluating text simplification, the aforementioned SARI, ROUGE, FRE, and FKGL, were not natural ways of evaluating a decomplication task and did not perform as well as we initially hoped for it to.

6.2 Future Work

Although there has been much progress in the field of text simplification, most metrics like SARI place emphasis on shortening the original sentence. On the contrary, we might instead want to use more tokens to shed more light on the original dense idea when decomplicating terms, and this notion is not yet captured on currently available metrics besides human evaluation.

Perhaps we could experiment with creating new metrics specifically designed for large language models to convert complex ideas into layman terms. Until then, we conclude that current metrics are unsatisfactory in our medical jargon use case.

In addition to knowledge distillation on the last layer attention modules, performing knowledge distillation on the first layer attention modules could help with preserving attention on the prompts. Furthermore, we could use a combination of both knowledge distillation and fine-tuning to help the model better learn the prompts. It is also possible that T5-Small is too small of a model for medical text translation and a T5-base model could be used as an alternative.

References

Chandrayee Basu, Rosni Vasu, Michihiro Yasunaga, and Qian Yang. 2023. Med-easi: Finely annotated dataset and models for controllable simplification of medical texts. *arXiv preprint arXiv:2302.09155*.

R. Hendawi, S. Alian, and J. Li. 2022. A smart mobile app to simplify medical documents and improve health literacy: System design and feasibility validation. *JMIR Formative Research*, 6(4):e35069.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

D. Kauchak and G. Leroy. 2016. Moving beyond readability metrics for health-related text simplification. *IT Professional*, 18(3):45–51. Epub 2016 May 25.

JP Kincaid, RP Fishburne, RL Rogers, and BS Chissom. 1975. Research branch report 8–75. *Memphis: Naval Air Station*.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

G. Leroy, D. Kauchak, and O. Mouradi. 2013. A user-study measuring the effects of lexical simplification and coherence enhancement on perceived and actual text difficulty. *International Journal of Medical Informatics*, 82(8):717–730. Epub 2013 Apr 29.

A. Phatak, D.W. Savage, R. Ohle, J. Smith, and V. Mago. 2022. Medical text simplification using reinforcement learning (teslea): Deep learning-based text simplification approach. *JMIR Medical Informatics*, 10(11):e38095.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788.