

Original Paper

# Medical Text Simplification Using Reinforcement Learning (TESLEA): Deep Learning–Based Text Simplification Approach

Atharva Phatak<sup>1</sup>, MSc; David W Savage<sup>2</sup>, MD, PhD; Robert Ohle<sup>3</sup>, MSc, MA, MBBCh; Jonathan Smith<sup>2</sup>, MD; Vijay Mago<sup>1</sup>, PhD

<sup>1</sup>Department of Computer Science, Lakehead University, Thunder Bay, ON, Canada

<sup>2</sup>NOSM University, Thunder Bay, ON, Canada

<sup>3</sup>NOSM University, Sudbury, ON, Canada

**Corresponding Author:**

Atharva Phatak, MSc  
Department of Computer Science  
Lakehead University  
955 Oliver Road  
Thunder Bay, ON, P7B 5E1  
Canada  
Phone: 1 8073558351  
Email: [phataka@lakeheadu.ca](mailto:phataka@lakeheadu.ca)

## Abstract

**Background:** In most cases, the abstracts of articles in the medical domain are publicly available. Although these are accessible by everyone, they are hard to comprehend for a wider audience due to the complex medical vocabulary. Thus, simplifying these complex abstracts is essential to make medical research accessible to the general public.

**Objective:** This study aims to develop a deep learning–based text simplification (TS) approach that converts complex medical text into a simpler version while maintaining the quality of the generated text.

**Methods:** A TS approach using reinforcement learning and transformer–based language models was developed. Relevance reward, Flesch-Kincaid reward, and lexical simplicity reward were optimized to help simplify jargon-dense complex medical paragraphs to their simpler versions while retaining the quality of the text. The model was trained using 3568 complex-simple medical paragraphs and evaluated on 480 paragraphs via the help of automated metrics and human annotation.

**Results:** The proposed method outperformed previous baselines on Flesch-Kincaid scores (11.84) and achieved comparable performance with other baselines when measured using ROUGE-1 (0.39), ROUGE-2 (0.11), and SARI scores (0.40). Manual evaluation showed that percentage agreement between human annotators was more than 70% when factors such as fluency, coherence, and adequacy were considered.

**Conclusions:** A unique medical TS approach is successfully developed that leverages reinforcement learning and accurately simplifies complex medical paragraphs, thereby increasing their readability. The proposed TS approach can be applied to automatically generate simplified text for complex medical text data, which would enhance the accessibility of biomedical research to a wider audience.

(*JMIR Med Inform* 2022;10(11):e38095) doi: [10.2196/38095](https://doi.org/10.2196/38095)

**KEYWORDS**

medical text simplification; reinforcement learning; natural language processing; manual evaluation

## Introduction

**Background**

Research from the field of biomedicine contains essential information about new clinical trials on topics related to new drugs and treatments for a variety of diseases. Although this

information is publicly available, it often has complex medical terminology, making it difficult for the general public to understand. One way to address this problem is by converting the complex medical text into a simpler language that can be understood by a wider audience. Although manual text simplification (TS) is one way to address the problem, it cannot be scaled to the rapidly expanding body of biomedical literature.

Therefore, there is a need for the development of *natural language processing* approaches that can automatically perform TS.

## Related Studies

### TS Approaches

Initial research in the field of TS focused on *lexical simplification* (LS) [1,2]. An LS system typically involves replacing complex words with their simpler alternatives using lexical databases, such as the *Paraphrase Database* [3], WordNet [4], or using language models, such as *bidirectional encoder representations from transformers* (BERT) [5]. Recent research defines TS as a *sequence-to-sequence* (seq2seq) task and has approached it by leveraging model architectures from other seq2seq tasks such as machine translation and text summarization [6-8]. Nisioi et al [9] proposed a neural *seq2seq* model, which used *long short-term memories* (LSTMs) for automatic TS. It was trained on simple-complex sentence pairs and showed through human evaluations that the TS system-generated outputs ultimately preserved meaning and were grammatically correct [9]. Afzal et al [10] incorporated LSTMs to create a quality-aware text summarization system for medical data. Zhang and Lapata [11] developed an LSTM-based neural encoder-decoder TS model and trained it using *reinforcement learning* (RL) to directly optimize SARI [12] scores along with a few other rewards. SARI is a widely used metric for automatic evaluation of TS.

With the recent progress in natural language processing research, LSTM-based models were outperformed by transformer [13]-based language models [13-16]. Transformers follow an encoder-decoder structure with both the encoder and decoder made up of  $L$  identical layers. Each layer consists of 2 sublayers, one being a feed-forward layer and the other a multihead attention layer. Transformer-based language models, such as BART [14], generative pretraining transformer (GPT) [15], and *text-to-text-transfer-transformer* [16], have achieved strong performance on natural language generation tasks such as text summarization and machine translation.

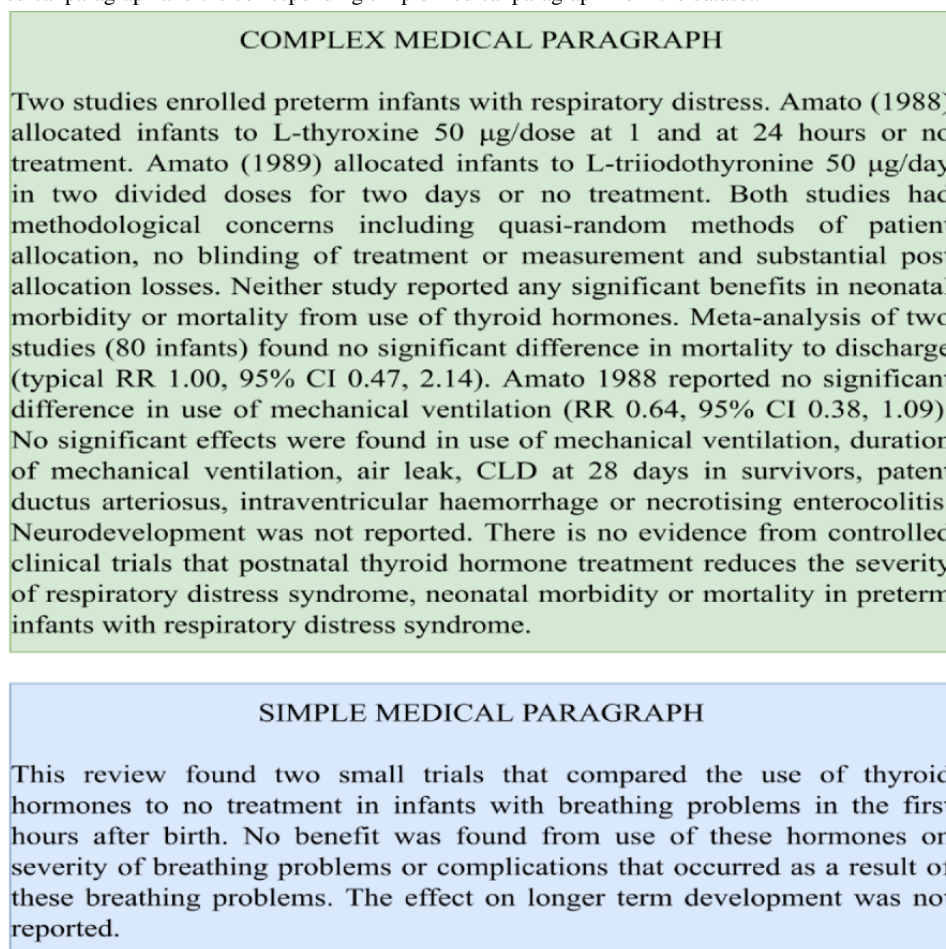
Building on the success of transformer-based language models, recently Martin et al [17] introduced *multilingual unsupervised sentence simplification* (MUSS) [17], a BART [14]-based language model, which achieved state-of-the-art performance on TS benchmarks by training on paraphrases mined from CCNet [18] corpus. Zhao et al [19] proposed a semisupervised approach that incorporated the back-translation architecture along with denoising autoencoders for the purpose of automatic TS. Unsupervised TS is also an active area of research but has

been primarily limited to LS. However, in a recent study, Surya et al [20] proposed an unsupervised approach to perform TS at both the lexical and syntactic levels. In general, research in the field of TS has been focused mostly on sentence-level simplification. However, Sun et al [21] proposed a document-level data set (D-wikipedia) and baseline models to perform document-level simplification. Similarly, Devaraj et al [8] proposed a BART [14]-based model that was trained using unlikelihood loss for the purpose of paragraph-level medical TS. Although their training regime penalizes the terms considered “jargon” and increases the readability, the generated text has lower quality and diversity [8]. Thus, the lack of document- or paragraph-level simplification makes this an important work in the advancement of the field.

### TS Data Sets

The majority of TS research uses data extracted from Wikipedia and news articles [11,22,23]. These data sets are paired sentence-level data sets (ie, for each complex sentence, there is a corresponding simple sentence). TS systems have heavily relied on sentence-level data sets, extracted from regular and simple English Wikipedia, such as WikiLarge [11], because they are publicly available. It was later shown by Xu [24] that there are issues with data quality for the data sets extracted from Wikipedia. They proposed the Newsela corpus, which was created by educators who rewrote news articles for different school-grade levels. Automatic sentence alignment methods [25] were used on the Newsela corpus to create a sentence-level TS data set. Despite the advancements in research on sentence-level simplification, there is a need for TS systems that can simplify text at a paragraph level.

Recent work has focused on the construction of document-level simplification data sets [17,21,26]. Sun et al [21] constructed a document-level data set, called D-Wikipedia, by aligning the English Wikipedia and Simple English Wikipedia spanning 143,546 article pairs. Although there are many data sets available for sentence-level TS, data sets for domain-specific paragraph-level TS are lacking. In the field of medical TS, Van den Bercken et al [27] constructed a sentence-level simplification data set using sentence alignment methods. Recently, Devaraj et al [8] proposed the first paragraph-level medical simplification data set, containing 4459 simple-complex pairs of text, and this is the data set used for the analysis and baseline training in this study. A snippet of a complex paragraph and its simplified version from the data set proposed by Devaraj et al [8] is shown in Figure 1. The data set is open sourced and publicly available [28].

**Figure 1.** Complex medical paragraph and the corresponding simple medical paragraph from the dataset.

### TS Evaluation

The evaluation of TS usually falls into 2 categories: automatic evaluations and manual (ie, human) evaluations. Because of the subjective nature of TS, it has been suggested that the best approach is to perform manual evaluations, based on criteria such as fluency, meaning preservation, and simplicity [20]. Automatic evaluation metrics most commonly used include readability indices such as Flesch-Kincaid Reading Ease [29], *Flesch-Kincaid Grade Level* (FKGL) [29], *Automated Readability Index* (ARI), Coleman-Liau index, and metrics for natural language generation tasks such as SARI [12] and BLEU [30].

Readability indices are used to assign a grade level to text signifying its simplicity. All the readability indices are calculated using some combination of word weighting, syllable, letter, or word counts, and are shown to measure some level of simplicity. Automatic evaluation metrics, such as BLEU [30] and SARI [12], are widely used in TS research, with SARI [12] having specifically been developed for TS tasks. SARI is computed by comparing the generated simplifications with both the source and target references. It computes an average of  $F_1$ -score for 3 *n-gram* overlap operations: additions, keeps, and deletions. Both BLEU [30] and SARI [12] are *n-gram*-based metrics, which may fail to capture the semantics of the generated text.

### Objective

The aim of this study is to develop an automatic TS approach that is capable of simplifying medical text data at a paragraph level, with the goal of providing greater accessibility of biomedical research. This paper uses RL-based training to directly optimize 2 properties of simplified text: relevance and simplicity. *Relevance* is defined as simplified text that retains salient and semantic information from the original article. *Simplicity* is defined as simplified text that is easy to understand and lexically simple. These 2 properties are optimized using TS-specific rewards, resulting in a system that outperforms previous baselines on Flesch-Kincaid scores. Extensive human evaluations are conducted with the help of domain experts to judge the quality of the generated text.

The remainder of the paper is organized as follows: The “Methods” section provides details on the data set, the training procedure, and the proposed model, and describes how automatic and human evaluations were conducted to analyze the outputs generated by the proposed model (TESLEA). The “Results” section provides a brief description of the baseline models and the results obtained by conducting automatic and manual evaluation of the generated text. Finally under the “Discussion” section, we highlight the limitations, future work, and draw conclusions.

## Methods

### Model Objective

Given a complex medical paragraph, the goal of this work is to generate a simplified paragraph that is concise and captures the salient information expressed in the complex text. To accomplish this, an RL-based simplification model is proposed, which optimizes multiple rewards during training, and is tuned using a paragraph-level medical TS data set.

### Data Set

The Cochrane Database of Scientific Reviews is a health care database with information on a wide range of clinical topics. Each review includes a plain language summary (PLS) written by the authors who follow guidelines to structure the summaries. PLSs are supposed to be clear, understandable, and accessible, especially for a general audience not familiar with the field of medicine. PLSs are highly heterogeneous in nature, and are not paired (ie, for every complex sentence there may not be a corresponding simpler version). However, Devaraj et al [8] used the Cochrane Database of Scientific Reviews data to produce a paired data set, which has 4459 pairs of complex-simple text, with each text containing less than 1024 tokens so that it can be fed into the BART [14] model for the purpose of TS. The pioneering data set developed by Devaraj et al [8] is used in this study for training the models and is publicly available [28].

### TESLEA: TS Using RL

#### Model and Rewards

The TS solution proposed for the task of simplifying complex medical text uses an RL-based simplification model, which optimizes multiple rewards (*relevance reward*, *Flesch-Kincaid Grade rewards*, and *lexical simplicity rewards*) to achieve a more complete and concise simplification. The following subsections introduce the computation of these rewards, along with the training procedure.

#### Relevance Reward

Relevance reward measures how well the semantics of the target text is captured in its simplified version. This is calculated by computing the cosine similarity between the target text embedding ( $E_T$ ) and the generated text embedding ( $E_G$ ). BioSentVec [31], a text embedding model trained on medical documents, is used to generate the text embeddings. The steps to calculate the relevance score are depicted in Algorithm 1.

**Algorithm 1:** Relevance reward

```

Input: T: Target text, G: Generated text, M: Embedding Model
Output:  $R_{cosine}$ : Relevance Reward
Variables:  $E_T$ : Target sentence embedding,  $E_G$ : Generated sentence embedding
1 Function RelevanceReward( $T, G, M$ )
   /* Compute sentence embedding for Target sentence. */
2  $E_T \leftarrow \text{ComputeEmbedding}(T, M)$ 
   /* Compute sentence embedding for generated sentence. */
3  $E_G \leftarrow \text{ComputeEmbedding}(G, M)$ 
   /* Compute Cosine Similarity. */
4  $R_{cosine} \leftarrow \frac{E_T \cdot E_G}{\|E_T\| \cdot \|E_G\|}$ 
5 return  $R_{cosine}$ 

```

The *RelevanceReward* function takes 3 arguments as input, namely, target text ( $T$ ), generated text ( $G$ ), and the embedding model ( $M$ ). The function *ComputeEmbedding* takes the input text and embedding model ( $M$ ) as input and generates the relevant text embedding. Finally, cosine similarity between generated text embedding ( $E_G$ ) and target text embedding ( $E_T$ ) is calculated to get the reward (Algorithm 1, line 4).

#### Flesch-Kincaid Grade Reward

FKGL refers to the grade level that must be attained to comprehend the presented information. A higher FKGL score indicates that the text is more complex, and a lower score indicates that the text is simpler. The FKGL for a text ( $S$ ) is calculated using equation 1 [29]:

$$\text{FKGL}(S) = 0.38 \times (\text{total words}/\text{total sentences}) + 1.8 \times (\text{total syllables}/\text{total words}) - (15.59) \quad (1)$$

The FKGL reward ( $R_{Flesch}$ ) is designed to reduce the complexity of generated text and is calculated as presented in Algorithm 2.

**Algorithm 2:** Flesch kincaid reward

```

Input: T: Target text, G: Generated text
Output:  $R_{Flesch}$ : Flesch Kincaid Reward
Variables:  $r(T)$ : Target text flesch kincaid grade level,  $r(G)$ : generated text flesch kincaid grade level.
1 Function FleschKincaidReward( $T, G$ )
2  $r(T) \leftarrow \text{FKGLScore}(T)$ 
3  $r(G) \leftarrow \text{FKGLScore}(G)$ 
4  $R_{Flesch} \leftarrow (r(T) - r(G))/r(T)$ 
5 return  $R_{Flesch}$ 

```

In Algorithm 2, the function *FleschKincaidReward* takes 2 arguments as inputs, namely, generated text ( $G$ ) and target text ( $T$ ). The *FKGLScore* function calculates the FKGL for the given text. Once the FKGL for  $T$  and  $G$  is calculated, the Flesch-Kincaid reward ( $R_{Flesch}$ ) is calculated as the relative difference between  $r(T)$  and  $r(G)$  (Algorithm 2, line 4), where  $r(T)$  and  $r(G)$  denote the FKGL of the target and generated text.

#### Lexical Simplicity Reward

Lexical simplicity is used to measure whether the words in the generated text ( $G$ ) are simpler than the words in the source text ( $S$ ). Laban et al [26] proposed a lexical simplicity reward that uses the correlation between word difficulty and word frequency [32]. As word frequency follows *zipf law*, Laban et al [26] used it to design the reward function, which involves calculating *zipf* frequency of newly inserted words, that is,  $Z(G - S)$ , and deleted words, that is,  $Z(S - G)$ . The lexical simplicity reward is defined in the same way as proposed by Laban et al [26] and is described in Algorithm 3. The analysis of the data set proposed by Devaraj et al [8] revealed that 87% of simple and complex pairs have a value of  $\Delta Z(S, G) \approx 0.4$ , where  $\Delta Z(S, G) = Z(G - S) - Z(S - G)$  is the difference between the *zipf* frequency of inserted words and deleted words, with the value of lexical reward ( $R_{lexical}$ ) scaled between 0 and 1.

In Algorithm 3, *LexicalSimplicityReward* requires the source text ( $S$ ) and the generated text ( $G$ ) as the inputs. Functions *ZIPFInserted* [25] and *ZIPFDeleted* [25] calculate the *zipf* frequency of newly inserted words and the deleted words. Finally, the lexical reward ( $R_{lexical}$ ) is calculated and normalized, as described in line 5.

**Algorithm 3:** Lexical simplicity reward

---

**Input:**  $S$ : Source Text,  $G$ : Generated Text  
**Output:**  $R_{lexical}$ : Lexical Simplicity Reward  
**Variables:**  $Z(G-S)$ : Zipf frequency of inserted words,  $Z(S-G)$ : Zipf frequency of deleted words,  $\Delta Z(S,G)$ : Difference between Zipf frequency of inserted and Zipf frequency of deleted words

```

1 Function LexicalSimplicityReward( $S,G$ )
  /* Compute Zipf frequency of inserted words. */
2  $Z(G-S) \leftarrow ZIPFInserted(G,S)$ 
  /* Compute Zipf frequency of deleted words. */
3  $Z(S-G) \leftarrow ZIPFDeleted(G,S)$ 
4  $\Delta Z(S,G) \leftarrow Z(G-S) - Z(S-G)$ 
5  $R_{lexical} \leftarrow 1 - \frac{\Delta Z(S,G) - 0.4}{0.4}$ 
6 return  $R_{lexical}$ 

```

---

**Training Procedure and Baseline Model****Pretrained BART**

The baseline language model used in this study for performing simplification was BART [14], which is a transformer based encoder-decoder model that was pretrained using a denoising objective function. The decoder part of the model is autoregressive in nature, making it more suitable for sentence-generation tasks. Furthermore, the BART model achieves strong performance on natural language generation tasks such as summarization, which was demonstrated on XSum [33] and CNN/Daily Mail [34] data sets. In this case, a version of BART fine-tuned on XSUM [33] data set is being used.

**Language Model Fine-tuning**

Transformer-based language models are pretrained on a large corpus of text and later fine-tuned on a downstream task by minimizing the maximum likelihood loss ( $Lml$ ) function [3]. Consider a paired data set  $C$ , where each instance consists of a source sentence containing  $n$  tokens  $x = \{x_1, \dots, x_n\}$  and target sequence containing  $m$  tokens  $y = \{y_1, \dots, y_m\}$ , then the  $Lml$  function is given in equation 2 with the computation described in Algorithm 4.

$$Lml = - \sum_{t=1}^m \log p_{\theta}(y_t | y_{<t}, x) \quad (2)$$

where  $\theta$  represents the model parameters and  $y_{<t}$  denotes preceding tokens before the position  $t$  [35].

**Algorithm 4:** Mle update

---

**Input:**  $D$ : Dictionary,  $\theta$ : Language Model  
**Output:**  $Lml$ : Maximum Likelihood Loss  
**Variables:**  $logits$ : Output of the model

```

1 Function MLEUpdate( $\theta, D$ )
  /* FORWARD function returns the output of the model. */
2  $logits \leftarrow FORWARD(\theta, D)$ 
  /* Calculating maximum likelihood loss using logits and D */
  /*
3  $Lml \leftarrow MLELoss(logits, D)$ 
4 return  $Lml$ 

```

---

However, the results obtained by minimizing  $Lml$  are not always optimal. There are 2 main reasons for the degradation of results.

The first is called “exposure bias” [36], which occurs when the model expects gold-standard data at each step of training, but does not receive appropriate supervision during testing, resulting in an accumulation of errors during prediction. The second is called “representation collapse” [37], which is a degradation of the pretrained language model representations during fine-tuning. Ranzato et al [36] avoided the problem of exposure bias by directly optimizing the specific discrete metric instead of minimizing the  $Lml$  with the help of an RL-based algorithm called REINFORCE [38]. A variant of REINFORCE [38] called Self-Critical Sequence Training [39] was used in this study to directly optimize certain rewards specifically designed for TS; more information on this is provided in the following subsection.

**Self-critical Sequence Training**

TS can be formulated as an RL problem, where the “agent” (language model) interacts with the environment to take “action” (next word prediction) based on a learned “policy” ( $p_{\theta}$ ) defined by model parameters  $\theta$  while observing some rewards ( $R$ ). In this work, BART [14] was used as the language model, and the REINFORCE [38] algorithm was used to learn an optimal policy that maximizes rewards. Specifically, REINFORCE was used with a baseline to stabilize the training procedure using an objective function ( $Lpg$ ) with a baseline reward  $b$  (equation 3):

$$Lpg = -(r(y^s) - b) \sum_{i=1}^n \log p_{\theta}(y_i^s | y_1^s, \dots, y_{i-1}^s, S) \quad (3)$$

where  $p_{\theta}(y_i^s | \dots)$  denotes the probability of the  $i$ th word conditioned on a previously generated sampled sequence by the model;  $r(y^s)$  denotes the reward computed for a sentence generated using sampling; denotes the source sentence, and  $n$  is the length of the generated sentence. Rewards are computed as a weighted sum of the relevance reward ( $R_{cosine}$ ),  $R_{Flesch}$ , and lexical simplicity reward ( $R_{lexical}$ ; Figure 2) and are given by:

$$r(y^s) = \alpha \cdot R_{cosine} + \beta \cdot R_{Flesch} + \delta \cdot R_{lexical} \quad (4)$$

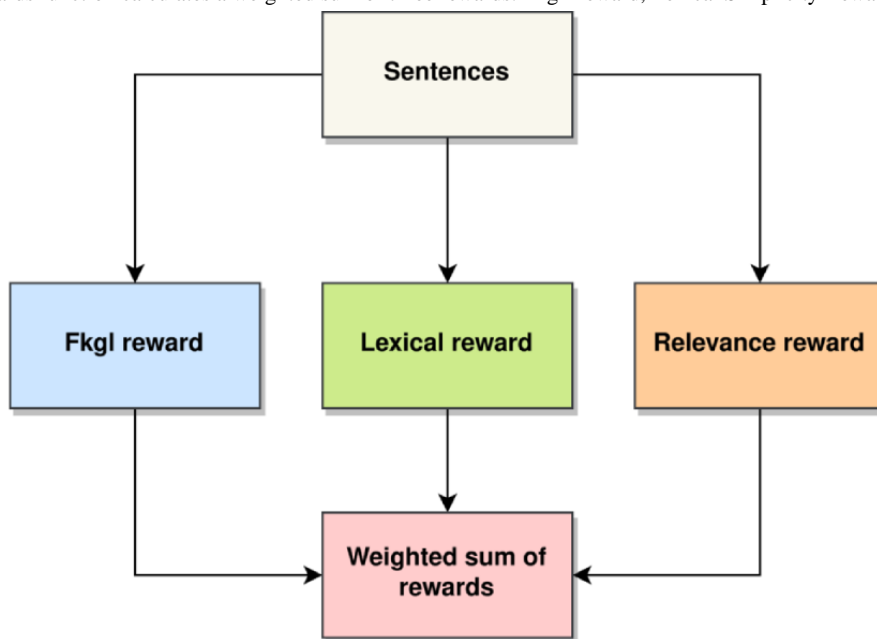
where  $\alpha$ ,  $\beta$ , and  $d$  are the weights associated with the rewards, respectively.

To approximate the baseline reward, Self-Critical Sequence Training [39] was used. The baseline was calculated by computing reward values for a sentence that has been generated using greedy decoding  $r(y^*)$  by the current model and its computation is described in Algorithm 5. The loss function is defined in equation 5:

$$Lpg = -(r(y^s) - r(y^*)) \sum_{i=1}^n \log p_{\theta}(y_i^s | y_1^s, \dots, y_{i-1}^s, S) \quad (5)$$

where  $y^*$  denotes the sentence generated using greedy decoding. More details on greedy decoding are described in [Multimedia Appendix 1](#) (see also [8,14,17,25,26,39-42]).

Figure 2. Compute Rewards function calculates a weighted sum of three rewards: Fkgl Reward, Lexical Simplicity Reward, Relevance Reward.



```

Algorithm 5: Self critical update


---


Input: D: Dictionary, M: Language Model
Output: Lpg : Policy Gradient Loss
Variables: ys: Sampled Sentence, y*: Greedy Sentence, n: length of
generated sequence, r(y*) : Reward for greedy sentence,
r(ys) : Reward for sampled sentence
Function SelfCriticalUpdate(θ, D)
  /* Generate sentence using multinomial Sampling. */
  ys ← GenerateSampleSentence(M, D)
  /* Generate sentence using Greedy Decoding. */
  y* ← GenerateGreedySentence(M, D)
  /* Compute reward for greedy sentence. */
  r(y*) ← ComputeRewards(y*, D)
  /* Compute reward for sampled sentence. */
  r(ys) ← ComputeRewards(ys, D)
  Lpg = (r(ys) - r(y*)) ∑i=1n log pθ(yis | y1s, ..., yi-1s)
  return Lpg
  
```

Intuitively, by minimizing the loss described in equation 5, the likelihood of choosing the samples sequence (y<sup>s</sup>) is promoted if the reward obtained for sampled sequence, r(y<sup>s</sup>), is greater than the reward obtained for the baseline rewards, that is, the samples that return higher reward than r(y<sup>\*</sup>). The samples that obtain a lower reward are subsequently suppressed. The model is trained using a combination of Lml and policy gradient loss similar to [43]. The overall loss is given as follows:

$$L = \gamma Lpg + (1 - \gamma) Lml \quad (6)$$

where  $\gamma$  is a scaling factor that can be tuned.

### Summary of the Training Process

Overall, the training procedure follows a 2-step approach. As the pretrained BART [14] was not trained on the medical domain-related text, it was first fine-tuned on the document-level paired data set [8] by minimizing the Lml (maximum likelihood estimation [MLE]; equation 2). In the

second part, the fine-tuned BART model was trained further using RL. The RL procedure of TESLEA involves 2 steps: (1) the RL step and (2) the MLE optimization step, which are both shown in Figure 3 and further described in Algorithm 6. The given simple-complex text pairs are converted to tokens as required by the BART model. In the MLE step, these tokens are used to compute logits from the model, and then finally MLE loss is computed. In the RL step, the model generates simplified text using 2 decoding strategies: (1) greedy decoding and (2) multinomial sampling. Rewards are computed as weighted sums (Figure 3) for sentences generated using both the decoding strategies. These rewards are then used to calculate the loss for the RL step. Finally, a weighted sum of losses is computed that is used to estimate the gradients and update model parameters. All the hyperparameter settings used are included in Multimedia Appendix 2 (see also [8,12,29,33,34,44-47]).

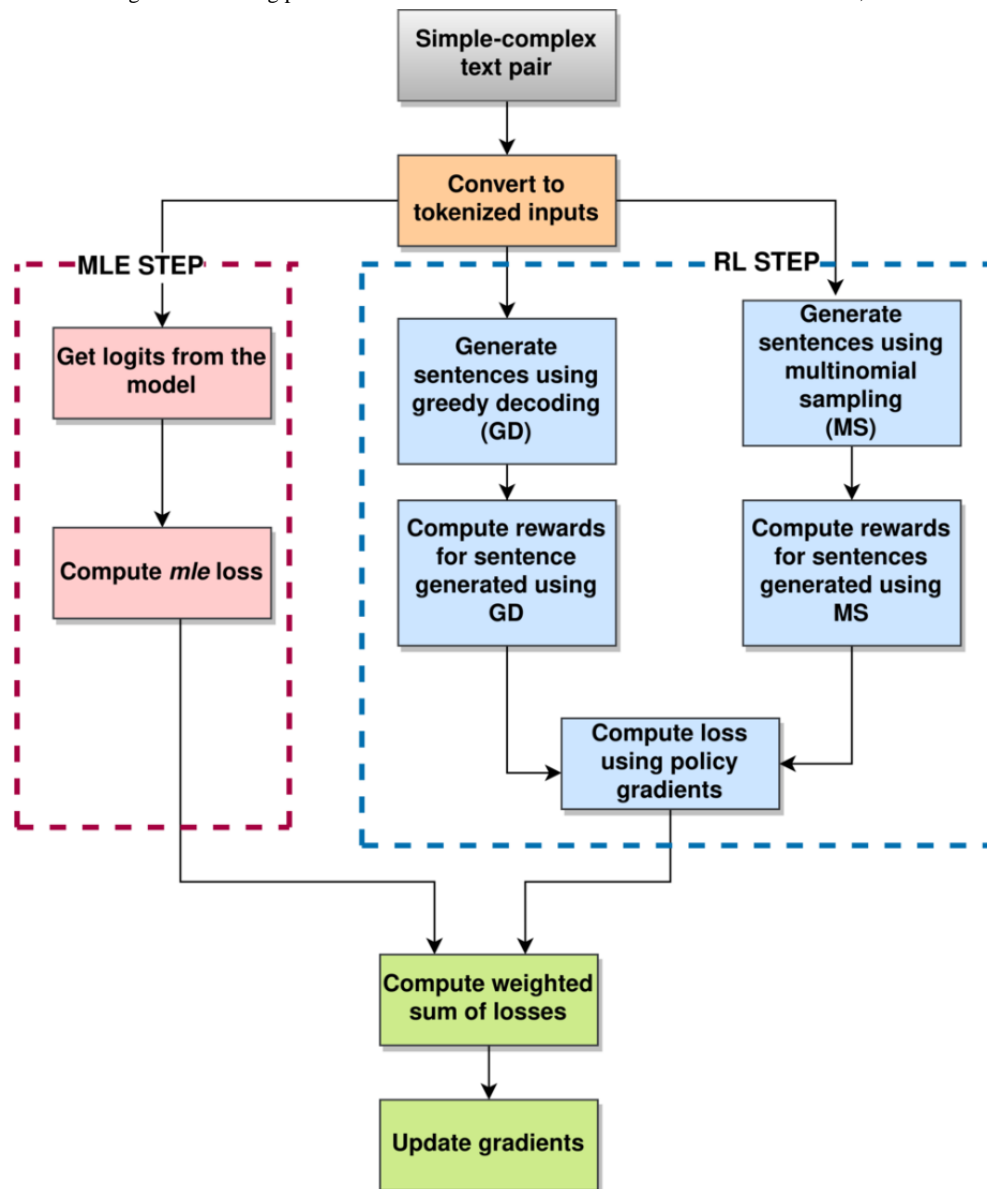
```

Algorithm 6: Training of simplification system


---


Input: Dpair : Paired Dataset, N: Iterations, γ : weight, M: Language
Model, Mf: Finetuned Language Model on paired
Dataset(Dpair)
Output: M: Language Model
1 M ← Mf
2 for i = 1 to N do
3   for batch ∈ Dpair do
4     D ← TOKENIZE(batch)
5     /* Calculate maximum likelihood loss. */
6     Lml ← MLEUpdate(M, D)
7     /* Calculate policy gradient loss. */
8     Lpg ← SelfCriticalUpdate(M, D)
9     /* Weighted sum of losses. */
10    L = γ · Lpg + (1 - γ) · Lml
11    Update model parameters with L
12  end
13 end
14 return Language Model θ
  
```

**Figure 3.** Reinforcement learning–based training procedure for TESLEA. MLE: maximum likelihood estimation; RL: reinforcement learning.



### Automatic Metrics

Two readability indices were used to perform automatic evaluations of the generated text, namely, FKGL and Automatic Readability Indices (ARIs). The SARI score is a standard metric for TS. The F-1 versions of ROUGE-1 and ROUGE-2 [44] scores were also reported. Readers can find more details about these metrics in [Multimedia Appendix 2](#). To measure the quality of the generated text, the criteria proposed by Yuan et al [45] were used, which are mentioned in the “Automatic Evaluation Metrics” section in [Multimedia Appendix 2](#). The criteria proposed by Yuan et al [45] can be automatically computed using a language model–based metric called “BARTScore.” Further details on how to use BARTScore to measure the quality of the generated text are also mentioned in [Multimedia Appendix 2](#).

### Human Evaluations

In this study, 3-domain experts judge the quality of the generated text based on the factors mentioned in the previous section. The evaluators rate the text on a Likert scale from 1 to 5. First, simplified test data were generated using TESLEA, and then 51 generated paragraphs were randomly selected, creating 3 subsets containing 17 paragraphs each. Every evaluator was presented with 2 subsets, that is, a total of 34 complex-simple TESLEA-generated paragraphs. The evaluations were conducted via Google Forms, and the human annotators were asked to measure the quality of simplification for informativeness (INFO), fluency (FLU), coherence (COH), factuality (FAC), and adequacy (ADE) ([Figure 4](#)). All the data collected were stored in CSV files for statistical analysis.

**Figure 4.** A sample question seen by the human annotator.

<p style="text-align: center;"><b>Complex Medical Paragraph</b></p> <p>A total of 38 studies involving 7843 children were included. Following educational intervention delivered to children, their parents or both, there was a significantly reduced risk of subsequent emergency department visits (RR 0.73, 95% CI 0.65 to 0.81, N = 3008) and hospital admissions (RR 0.79, 95% CI 0.69 to 0.92, N = 4019) compared with control. There were also fewer unscheduled doctor visits (RR 0.68, 95% CI 0.57 to 0.81, N = 1009). Very few data were available for other outcomes (FEV1, PEF, rescue medication use, quality of life or symptoms) and there was no statistically significant difference between education and control. Asthma education aimed at children and their carers who present to the emergency department for acute exacerbations can result in lower risk of future emergency department presentation and hospital admission. There remains uncertainty as to the long-term effect of education on other markers of asthma morbidity such as quality of life, symptoms and lung function. It remains unclear as to what type, duration and intensity of educational packages are the most effective in reducing acute care utilisation.</p>	<p style="text-align: center;"><b>Generated Simple Medical Paragraph</b></p> <p>This review of studies found that education aimed at children and their carers reduces the need for future emergency department visits for acute exacerbations in children aged four to 16 years who suffer an asthma attack. Although education programmes have been effective at reducing the emergency department visit, there is uncertainty as to whether education programmes can have a long-term impact on other markers of asthma morbidity, such as quality of life, symptoms and breathing patterns.</p>
--	---

#### QUESTIONS

- Rate the Generated text on a scale to 1 to 5 considering the Informativeness
  1. No relevant information is retained in generated text
  2. Partial relevant information is retained in generated text
  3. Neutral/ Undecided
  4. Significant relevant information is retained in generated text
  5. All relevant information is retained in generated text
- Rate the Generated text on a scale to 1 to 5 considering the Fluency
  1. Fluency is lost in the generated text
  2. Fluency is partially lost in the generated text
  3. Neutral/ Undecided
  4. Fluency is partially maintained in the generated text.
  5. Fluency is maintained in the generated text.
- Rate the Generated text on a scale to 1 to 5 considering the Coherence
  1. Coherence is lost in the generated text.
  2. Coherence is partially lost in the generated text.
  3. Neutral/ Undecided.
  4. Coherence is partially maintained in the generated text.
  5. Coherence is maintained in the generated text.
- Rate the Generated text on a scale to 1 to 5 considering the Factuality
  1. Factuality is lost in the generated text
  2. Factuality is partially lost in the generated text.
  3. Neutral/ Undecided
  4. Factuality is partially maintained in the generated text.
  5. Factuality is maintained in the generated text.
- Rate the Generated text on a scale to 1 to 5 considering the Adequacy.
  1. Adequacy is lost in the generated text
  2. Adequacy is partially lost in the generated text
  3. Neutral/ Undecided
  4. Adequacy is partially maintained in the generated text
  5. Adequacy is maintained in the generated text.

## Results

### Overview

This section consists of 3 subsections, namely, (1) Baseline Models, (2) Automatic Evaluations, and (3) Human Evaluations. The first section highlights the baseline models used for comparison and analysis. The second section discusses the results obtained by performing automatic evaluations of the

model. The third and final section discusses the results obtained from human assessments and analyzes the relationship between human annotations and automatic metrics.

### Baseline Models

TESLEA is compared with other strong baseline models and their details are discussed below:

- **BART-Fine-tuned:** BART-Fine-tuned is a BART-large model fine-tuned using an *Lml* on the data set proposed by



- Devaraj et al [8]. Studies have shown that large pretrained models often perform competitively when fine-tuned for downstream tasks, thus making this a strong competitor.
- BART-UL: Devaraj et al [8] also proposed BART-UL for paragraph-level medical TS. It is the first model to perform paragraph-level medical TS and has achieved strong results on automated metrics. BART-UL was trained using an unlikelihood objective function that penalizes the model for generating technical words (ie, complex words). Further details on the training procedure of BART-UL are described in [Multimedia Appendix 1](#).
  - MUSS: MUSS [17] is a BART-based language model that was trained by mining paraphrases from the CCNet corpus [18]. MUSS was trained on a data set consisting of 1 million paraphrases, helping it achieve a strong SARI score. Although MUSS is trained on a sentence-level data set, it still serves as a strong baseline for comparison. Further details on the training procedure for MUSS are discussed in [Multimedia Appendix 1](#).
  - Keep it Simple (KIS): Laban et al [26] proposed an unsupervised approach for paragraph-level TS. KIS is trained using RL and uses the GPT-2 model as a backbone. KIS has shown strong performance on SARI scores beating many supervised and unsupervised TS approaches.
- Additional details on the training procedure for KIS are described in [Multimedia Appendix 1](#).
- PEGASUS models: PEGASUS is a transformer-based encoder-decoder model that has achieved state-of-the-art results on many text-summarization data sets. It was specifically designed for the task of text summarization. In our analysis, we used 2 variants of PEGASUS models, namely, (1) PEGASUS-large, the large variant of Pegasus model, and (2) PEGASUS-pubmed-large, the large variant of the PEGASUS model that was pretrained on a PubMed data set. Both the PEGASUS models were fine-tuned using *Lml* on the data set proposed by Devaraj et al [8]. For more information regarding the PEGASUS model, the readers are suggested to refer to [46].

The models described above are the only ones available for medical TS as of June 2022.

### Results of Automatic Metrics

The metrics used for automatic evaluation are FKGL, ARI, ROUGE-1, ROUGE-2, SARI, and BARTScore. The mean readability indices scores (ie, FKGL and ARI) obtained by various models are reported in [Table 1](#). ROUGE-1, ROUGE-2, and SARI scores are reported in [Table 2](#) and BARTScore is reported in [Table 3](#).

**Table 1.** Flesch-Kincaid Grade Level and Automatic Readability Index for the generated text.<sup>a</sup>

Text	Flesch-Kincaid Grade Level	Automatic Readability Index
<b>Baseline</b>		
Technical abstracts	14.42	15.58
Gold-standard references	13.11	15.08
<b>Model generated</b>		
BART-Fine-tuned	13.45	15.32
BART-UL	11.97	13.73 <sup>b</sup>
TESLEA	11.84 <sup>b</sup>	13.82
MUSS <sup>c</sup>	14.29	17.29
Keep it Simple	14.15	17.05
PEGASUS-large	14.53	17.55
PEGASUS-pubmed-large	16.35	19.8

<sup>a</sup>TESLEA significantly reduces FKGL and ARI scores when compared with plain language summaries.

<sup>b</sup>Best score.

<sup>c</sup>MUSS: multilingual unsupervised sentence simplification.

**Table 2.** ROUGE-1, ROUGE-2, and SARI scores for the generated text.<sup>a</sup>

Model	ROUGE-1	ROUGE-2	SARI
BART-Fine-tuned	0.40	0.11	0.39
BART-UL	0.38	0.14	0.40 <sup>b</sup>
TESLEA	0.39	0.11	0.40 <sup>b</sup>
MUSS <sup>c</sup>	0.23	0.03	0.34
Keep it Simple	0.23	0.03	0.32
PEGASUS-large	0.44 <sup>b</sup>	0.18 <sup>b</sup>	0.40 <sup>b</sup>
PEGASUS-pubmed-large	0.42	0.16	0.40 <sup>b</sup>

<sup>a</sup>TESLEA achieves similar performance to other models. Higher scores of ROUGE-1, ROUGE-2, and SARI are desirable.

<sup>b</sup>Best performance.

<sup>c</sup>MUSS: multilingual unsupervised sentence simplification.

**Table 3.** Faithfulness Score and F-score for the generated text by the models.<sup>a</sup>

Models	Faithfulness Score	F-score
BART-Fine-tuned	0.137	0.078
BART-UL	0.242	0.061
TESLEA	0.366 <sup>b</sup>	0.097 <sup>b</sup>
MUSS <sup>c</sup>	0.031	0.029
Keep it Simple	0.030	0.028
PEGASUS-large	0.197	0.073
PEGASUS-pubmed-large	0.29	0.063

<sup>a</sup>Higher scores of Faithfulness and F-score are desirable.

<sup>b</sup>Highest score.

<sup>c</sup>MUSS: multilingual unsupervised sentence simplification.

### Readability Indices, ROUGE, and SARI Scores

The readability indices scores reported in [Table 1](#) suggest that the FKGL scores obtained by TESLEA are better (ie, a lower score) when compared with the FKGL scores obtained by comparing technical abstracts (ie, complex medical paragraphs available in the data set) with the gold-standard references (ie, simple medical paragraphs corresponding to the complex medical paragraphs). Moreover, TESLEA achieves the lowest FKGL score (11.84) when compared with baseline models, indicating significant improvement in the TS. The results suggest that (1) BART-based transformer models are capable of performing simplification at the paragraph level such that the outputs are at a reduced reading level (FKGL) when compared with technical abstracts, gold-standard references, and baseline models. (2) The proposed method to optimize TS-specific rewards allows the generation of text with greater readability than even the gold-standard references, as indicated by the FKGL scores in [Table 1](#). The reduction in FKGL scores can be explained by the fact that FKGL was a part of a reward ( $R_{Flesch}$ ) that was directly being optimized.

In addition, we report the SARI [12] and ROUGE scores [44] as shown in [Table 2](#). SARI is a standard automatic metric used

in sentence-level TS tasks. The ROUGE score is another standard metric in text summarization tasks. The results show that TESLEA matches the performance of baseline models on both ROUGE and SARI scores. Although there are no clear patterns when ROUGE and SARI scores are considered, there are differences in the quality of text generated by these models and these are explained in the “Text Quality Measure” subsection.

### Text Quality Measure

There has been significant progress in designing automatic metrics that are able to capture linguistic quality of the text generated by language models. One such metric that is able to measure the quality of generated text is BARTScore [45]. BARTScore has shown strong correlation with human assessments on various tasks ranging from machine translation to text summarization. BARTScore has 4 different metrics (ie, Faithfulness Score, Precision, Recall, F-score), which can be used to measure different qualities of generated text. Further details on how to use BARTScore are mentioned in [Multimedia Appendix 2](#).

According to the analysis conducted by Yuan et al [45], Faithfulness Score measures 3 aspects of generated text via

COH, FLU, and FAC. The F-score measures 2 aspects of generated text (INFO and ADE). In our analysis, we use these 2 variants of BARTScore to measure COH, FLU, FAC, INFO, and ADE. TESLEA achieves the highest values (Table 3) of Faithfulness Score (0.366) and F-score (0.097), indicating that the rewards designed for the purpose of TS not only help the model in generating simplified text but also on some level preserve the quality of generated text. The F-scores of all the models are relatively poor (ie, scores closer to 1 are desirable). One of the reasons for low F-scores could be the introduction of misinformation or hallucinations in the generated text, a common problem for language models, which could be addressed by adapting training strategies that focus on INFO via the help of rewards or objective functions.

For qualitative analysis we randomly selected 50 sentences from the test data and calculated the average number of tokens based on BART model vocabulary. For the readability measure, we calculated the FKGL scores of these generated texts and noted any textual inconsistencies such as misinformation. The analysis revealed that the text generated by most models was significantly smaller than the gold-standard references (Table 4). Furthermore, TESLEA- and BART-UL-generated texts were significantly shorter compared with other baseline models and TESLEA had the lowest FKGL score among all the models as depicted in Table 4.

From a qualitative point of view, the sentences generated by most baseline models involve significant duplication of text from the original complex medical paragraph. The outputs

generated by the KIS model were incomplete and appear “noisy” in nature. One of the reasons for the noise generation could be because of unstable training due to lack of a huge corpus of domain-specific data. BART-UL-generated paragraphs are simplified as indicated by the FKGL and ARI scores, but they are extractive in nature (ie, the model learns to select simplified sentences from the original medical paragraph and combines them to form a simplification). PEGASUS-pubmed-large-generated paragraphs are also extractive in nature and similar to BART-UL-generated paragraphs, but it was observed that they were grammatically inconsistent. In contrast to baseline models, the text generated by TESLEA was concise, semantically relevant, and simple, without involving any medical domain-related complex vocabulary. Figure 5 shows an example of text generated by all the models, with blue text indicating the copied text.

In addition to the duplicated text, the models also induced misinformation in the generated text. The most common form of induced misinformation observed was “The evidence is current up to [date],” as shown in Figure 6. This text error occurred due to the structure of the data (ie, PLS contains statements related to this research, but these statements were not in the original text; thus, the model attempted to add these statements to the generated text although it is not factually correct). Thus considerable attention should be paid to including FAC measures in the training regime of these models. For a more complete assessment of the quality of simplification, human evaluation was conducted using domain experts for the text generated by TESLEA.

**Table 4.** Average number of tokens and average Flesch-Kincaid Grade Level scores for selected samples.

Model	Number of tokens	Flesch-Kincaid Grade Level
Technical abstracts	498.11	14.37
Gold-standard references	269.74	12.77
TESLEA	131.37	12.34
BART-UL	145.08	12.66
Keep it Simple	187.59	13.78
Multilingual unsupervised sentence simplification	193.07	13.86
PEGASUS-large	272.04	13.93
PEGASUS-pubmed-large	150.00	15.09

**Figure 5.** Comparison of Text Generated by all the models. The highlighted blue text indicates copying. CI: Confidence Interval; FEV: Force Expiratory Volume; N: Population size; PEV: Peak Expiratory Flow; RR: Respiratory Rate.

<p style="text-align: center;"><b>COMPLEX MEDICAL PARAGRAPH</b></p> <p>A total of 38 studies involving 7843 children were included. Following educational intervention delivered to children, their parents or both, there was a significantly reduced risk of subsequent emergency department visits (RR 0.73, 95% CI 0.65 to 0.81, N = 3008) and hospital admissions (RR 0.79, 95% CI 0.69 to 0.92, N = 4019) compared with control. There were also fewer unscheduled doctor visits (RR 0.68, 95% CI 0.57 to 0.81, N = 1009). Very few data were available for other outcomes (FEV1, PEF, rescue medication use, quality of life or symptoms) and there was no statistically significant difference between education and control. Asthma education aimed at children and their carers who present to the emergency department for acute exacerbations can result in lower risk of future emergency department presentation and hospital admission. There remains uncertainty as to the long-term effect of education on other markers of asthma morbidity such as quality of life, symptoms and lung function. It remains unclear as to what type, duration and intensity of educational packages are the most effective in reducing acute care utilisation.</p>
<p style="text-align: center;"><b>BART-FINETUNED GENERATED PARAGRAPH</b></p> <p>A total of 38 studies involving 7843 children were included in this review. Following educational intervention delivered to children, their parents or both, there was a significantly reduced risk of subsequent emergency department visits (RR 0.73, 95% CI 0.65 to 0.81, N = 3008) and hospital admissions compared with control. There were also fewer unscheduled doctor visits. There remains uncertainty as to the long-term effect of education on other markers of asthma morbidity such as quality of life, symptoms and lung function.</p>
<p style="text-align: center;"><b>KIS GENERATED PARAGRAPH</b></p> <p>A total of 38 studies involving 7843 children were included. Following educational intervention, parents or both, there was a significant reduction in the risk of emergency department visits for children who had experienced trouble at school. RR 0.73, the most common estimate of the risk of emergency department visits, was also found to be associated with asthma symptoms (RR 0.65, 95% CI 0.65 to 0.81, N = 3008).</p>
<p style="text-align: center;"><b>PL GENERATED PARAGRAPH</b></p> <p>This review of 38 studies involving 7843 children found that asthma education delivered to children, their parents or both, can reduce the risk of future emergency department visits and hospital admissions. There were also fewer unscheduled doctor visits. Very few data were available for other outcomes (FEV1, PEF, rescue medication use, quality of life or symptoms) and there was no statistically significant difference between education and control. Asthma education aimed at children and their carers who present to the emergency department for acute exacerbations can result in lower risk of future emergency department presentation and hospital admission. There remains uncertainty as to the long-term effect of education on other markers of asthma morbidity such as quality of life, symptoms and lung function. It remains unclear as to what type, duration and intensity of educational packages are the most effective in reducing acute care utilisation.</p>
<p style="text-align: center;"><b>PPL GENERATED PARAGRAPH</b></p> <p>We review found that education aimed at children and their carers who present to the emergency department for acute exacerbations can result in a lower risk of future emergency department visits and hospital admission. We remains uncertain as to the long-term effect of education on other markers of asthma morbidity such as quality of life, symptoms and lung function.</p>
<p style="text-align: center;"><b>MUSS GENERATED PARAGRAPH</b></p> <p>A total of 38 studies involving 7843 children were included. Following educational intervention delivered to children, their parents or both, there was a significantly reduced risk of subsequent emergency department visits (RR 0.73, 95% CI 0.65 to 0.81, N = 3008) and hospital admissions (RR 0.79, 95% CI 0.69 to 0.92, N = 4019). There were also fewer unscheduled doctor visits (RR 0.68, 95% CI 0.57 to 0.81, N = 1009). There was no statistically significant difference between education and control. Asthma education aimed at children and their carers who present to the emergency department for acute exacerbations can result in lower risk of future emergency department presentation and hospital admission. It remains unclear as to what type, duration and intensity of educational packages are the most effective in reducing acute care utilisation.</p>
<p style="text-align: center;"><b>BART-UL GENERATED PARAGRAPH</b></p> <p>This systematic review identified 38 studies involving 7843 children. These studies found that asthma education aimed at children and their carers who present to the emergency department for acute exacerbations can result in lower risk of future emergency department presentation and hospital admission. However, there is uncertainty as to what type, duration and intensity of educational packages are the most effective in reducing acute care utilisation. There remains uncertainty about the long-term effect of education on other markers of asthma morbidity such as quality of life, symptoms and lung function.</p>
<p style="text-align: center;"><b>TESLEA GENERATED PARAGRAPH</b></p> <p>This review of studies found that education aimed at children and their carers reduces the need for future emergency department visits for acute exacerbations in children who suffer an asthma attack. Although education programmes have been effective at reducing the emergency department visits, there is uncertainty as to whether education programmes can have a long-term impact on other markers of asthma morbidity, such as quality of life, symptoms and breathing patterns.</p>

**Figure 6.** Example of misinformation found in Generated text. CIDSL: Cornelia de Lange syndrome; IVIg: Intravenous immune globulin; MS: Multiple Sclerosis; PE: plasma exchange.

**Generated Text**

Twelve trials including a total of 1211 trials were included in this review. Seven trials compared IVIg with PE and compared it with PE. **The evidence is current up to July 2013.** These trials were from all over the world and include people with CIDSL and MS and include people with and without MS from all walks of life. The findings of this review suggest that, in severe cases of MS, IVIg, given within two weeks of onset of the disease, hastens recovery as much as PE therapy.

## Human Evaluations

For this research, 3 domain experts assessed the quality of generated text, based on factors of INFO, FLU, COH, FAC, and ADE, as proposed by Yuan et al [45], which are discussed in [Multimedia Appendix 2](#). To measure interrater reliability, the percentage agreement between the annotators is calculated, and the results are shown in [Table 5](#). The average percentage agreement for the factors of FLU, COH, FAC, and ADE is the highest, indicating that annotators agree among their evaluations.

The average Likert score for each factor is also reported by each rater ([Table 6](#)). From the data mentioned in [Table 6](#), the raters

think that the COH and FLU have the highest quality, with the ADE, FAC, and INFO also rated reasonably high.

To further assess whether results obtained by automated metrics truly signify an improvement in the quality of generated text by TESLEA, the Spearman rank correlation coefficient was calculated between human ratings and the automatic metrics for all 51 generated paragraphs (text), with the results shown in [Table 7](#). The BARTScore has the highest correlation with human ratings for FLU, FAC, COH, and ADE compared with other metrics. A few text samples along with their human annotations and automated metric scores are shown in [Multimedia Appendix 3](#) and [Figure 7](#).

**Table 5.** Average percentage interrater agreement.

Interrater agreement	Informativeness, %	Fluency, %	Factuality, %	Coherence, %	Adequacy, %
A1 <sup>a</sup> and A2 <sup>b</sup>	82.35	82.35	82.35	70.59	82.35
A1 and A3 <sup>c</sup>	70.59	58.82	70.59	70.59	70.59
A3 and A2	52.94	70.59	74.51	74.51	64.71
Average (% agreement)	68.63	70.59	74.51	74.51	72.55

<sup>a</sup>A1: annotator 1.

<sup>b</sup>A2: annotator 2.

<sup>c</sup>A3: annotator 3.

**Table 6.** Average Likert score by each rater for informativeness, fluency, factuality, coherence, and adequacy.

Rater	Informativeness	Fluency	Factuality	Coherence	Adequacy
A1	3.82	4.12	3.91	3.97	3.76
A2	3.50	4.97	3.59	4.82	3.68
A3	4.06	3.94	3.85	3.94	3.85
Average Likert score	3.79	4.34	3.78	4.24	3.76

**Table 7.** Spearman rank correlation coefficient between automatic metrics and human ratings for the text generated by TESLEA.

Metric	Informativeness	Fluency	Factuality	Coherence	Adequacy
ROUGE-1	0.18 <sup>a</sup>	-0.04	-0.01	-0.05	0.06
ROUGE-2	0.08	-0.01	-0.05	-0.04	0.05
SARI	0.09	-0.66	-0.13	-0.01	0.01
BARTScore	0.08	0.32 <sup>a</sup>	0.38 <sup>a</sup>	0.22 <sup>a</sup>	0.07 <sup>a</sup>

<sup>a</sup>Best result.

Figure 7. Samples of Complex, Simple (Gold) and generated medical paragraphs along with automated metrics and Human annotations.



## Discussion

### Principal Findings

The most up-to-date research about biomedicine is often inaccessible to the general public due to the domain-specific medical terminology. A way to address this problem is by creating a system that converts complex medical information into a simpler form, thus making it accessible to everyone. In

this study, a TS approach was developed that can automatically simplify complex medical paragraphs while maintaining the quality of the generated text. The proposed approach trains the transformer-based BART model to optimize rewards specific for TS, resulting in increased simplicity. The BART model is trained using the proposed RL method to optimize certain rewards that help generate simpler text while maintaining the quality of generated text. As a result, the trained model generates simplified text that reduces the complexity of the original text

by 2-grade points, when measured using the FKGL [29]. From the results obtained, it can be concluded that TESLEA is effective in generating simpler text compared with technical abstracts, the gold-standard references (ie, simple medical paragraphs corresponding to complex medical paragraphs), and the baseline models. Although previous work [8] developed baseline models for this task, to the best of our knowledge, this is the first time RL is being applied to the field of medical TS. Moreover, previous studies failed to analyze the quality of the generated text, which this study measures via the factors of FLU, FAC, COH, ADE, and INFO. Manual evaluations of TESLEA-generated text were conducted with the help of domain experts using the aforesaid factors and further research was conducted to analyze which automatic metrics agree with manual annotations using the Spearman rank correlation coefficient. The analysis revealed that BARTScore [45] best correlates with the human annotations when evaluated for a text generated by TESLEA, indicating that TESLEA learns to generate semantically relevant and fluent text, which conveys the essential information mentioned in the complex medical paragraph. These results suggest that (1) TESLEA can perform TS of medical paragraphs such that outputs are simple and maintain the quality, (2) the rewards optimized by TESLEA help the model capture syntactic and semantic information, increasing the FLU and COH of outputs, as witnessed when the outputs are evaluated by BARTScore and human annotators.

### Limitations and Future Work

Although this research is a significant contribution to the literature on medical TS, the proposed approach does have a

few limitations, addressing which can result in even better outputs. TESLEA can generate simpler versions of the text, but in some instances, it induces misinformation, resulting in reduced FAC and INFO of the generated text. Therefore, there is a need to design rewards that consider the FAC and INFO of the generated text. We also plan to conduct extensive human evaluations on a large scale for the text generated by various models (eg, KIS, BART-UL) using domain experts (ie, physicians and medical students).

Transformer-based language models are sensitive to the pretraining regime, so a possible next step is to pretrain a language model on domain-specific raw data sets such as PubMed [40], which will help develop domain-specific vocabulary for the model. Including these strategies may help in increasing the simplicity of the generated text.

### Conclusion

The interest in and need for TS in the medical domain are of growing interest as the quantity of data is continuously increasing. Automated systems, such as the one proposed in this paper, can dramatically increase accessibility to information for the general public. This work not only provides a technical solution for automated TS, but also lays out and addresses the challenges of evaluating the outputs of such systems, which can be highly subjective. It is the authors' sincere hope that this work allows other researchers to build on and improve the quality of similar effort.

---

### Acknowledgments

The authors thank the research team at DaTALab, Lakehead University, for their support. The authors also thank Compute Canada for providing the computational resources without which this research would not have been possible. This research is funded by NSERC Discovery (RGPIN-2017-05377) held by Dr. Vijay Mago. The authors thank Mr. Aditya Singhal (MSc student at Lakehead University) for providing his feedback on the manuscript.

---

### Conflicts of Interest

None declared.

---

### Multimedia Appendix 1

Training Procedures and Decoding Methods.

[\[DOCX File , 129 KB-Multimedia Appendix 1\]](#)

---

### Multimedia Appendix 2

Hyperparameters and Evaluation Metrics.

[\[DOCX File , 190 KB-Multimedia Appendix 2\]](#)

---

### Multimedia Appendix 3

Abbreviations and Examples.

[\[DOCX File , 1060 KB-Multimedia Appendix 3\]](#)

---

### References

1. Carroll J, Minnen G, Pearce D, Canning Y, Devlin S, Tait J. Simplifying text for language-impaired readers. New Brunswick, NJ: Association for Computational Linguistics; 1999 Presented at: Ninth Conference of the European Chapter of the

- Association for Computational Linguistics; June 8-12, 1999; Bergen, Norway p. 269-270 URL: [https://aclanthology.org/E\[199-1042](https://aclanthology.org/E[199-1042)
2. Paetzold G, Specia L. Unsupervised Lexical Simplification for Non-Native Speakers. AAAI 2016 Mar 05;30(1):3761-3767 [FREE Full text] [doi: [10.1609/aaai.v30i1.9885](https://doi.org/10.1609/aaai.v30i1.9885)]
  3. Ganitkevitch J, Van Durme B, Callison-Burch C. PPDB: The paraphrase database. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Brunswick, NJ: Association for Computational Linguistics; 2013 Jun Presented at: The 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 9-12, 2013; Atlanta, GA p. 758-764 URL: <https://aclanthology.org/N13-1092> [doi: [10.3115/v1/p15-2070](https://doi.org/10.3115/v1/p15-2070)]
  4. Rebecca Thomas S, Anderson S. WordNet-Based Lexical Simplification of a Document. In: Proceedings of the 11th Conference on Natural Language Processing (KONVENS 2012). 2012 Presented at: The 11th Conference on Natural Language Processing (KONVENS 2012); September 19-21, 2012; Vienna, Austria p. 80 URL: [https://www.researchgate.net/publication/270450791\\_WordNet-Based\\_Lexical\\_Simplification\\_of\\_a\\_Document](https://www.researchgate.net/publication/270450791_WordNet-Based_Lexical_Simplification_of_a_Document)
  5. Qiang J, Li Y, Zhu Y, Yuan Y, Wu X. Lexical Simplification with Pretrained Encoders. AAAI 2020 Apr 03;34(05):8649-8656. [doi: [10.1609/aaai.v34i05.6389](https://doi.org/10.1609/aaai.v34i05.6389)]
  6. Zhu Z, Bernhard D, Gurevych I. A monolingual tree-based translation model for sentence simplification. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010). Beijing, China: Coling 2010 Organizing Committee; 2010 Presented at: The 23rd International Conference on Computational Linguistics (Coling 2010); August 23-27, 2010; Beijing, China p. 1353-1361 URL: <https://aclanthology.org/C10-1152.pdf>
  7. Wubben S, van den Bosch A, Kraemer E. Sentence simplification by monolingual machine translation. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). New Brunswick, NJ: Association for Computational Linguistics; 2012 Presented at: The 50th Annual Meeting of the Association for Computational Linguistics; July 8-14, 2012; Jeju Island, Korea p. 1015-1024 URL: <https://aclanthology.org/P12-1107>
  8. Devaraj A, Marshall I, Wallace B, Li J. Paragraph-level Simplification of Medical Texts. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Brunswick, NJ: Association for Computational Linguistics; 2021 Jun Presented at: The 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 6-11, 2021; Virtual p. 4972-4984 URL: <https://aclanthology.org/2021.naacl-main.395.pdf> [doi: [10.18653/v1/2021.naacl-main.395](https://doi.org/10.18653/v1/2021.naacl-main.395)]
  9. Nisioi S, Štajner S, Paolo Ponzetto S, Dinu LP. Exploring neural text simplification models. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). New Brunswick, NJ: Association for Computational Linguistics; 2017 Presented at: The 55th Annual Meeting of the Association for Computational Linguistics; July 30-August 4, 2017; Vancouver, BC p. 85-91 URL: <https://aclanthology.org/P17-2014.pdf> [doi: [10.18653/v1/p17-2014](https://doi.org/10.18653/v1/p17-2014)]
  10. Afzal M, Alam F, Malik KM, Malik GM. Clinical Context-Aware Biomedical Text Summarization Using Deep Neural Network: Model Development and Validation. J Med Internet Res 2020 Oct 23;22(10):e19810 [FREE Full text] [doi: [10.2196/19810](https://doi.org/10.2196/19810)] [Medline: [33095174](https://pubmed.ncbi.nlm.nih.gov/33095174/)]
  11. Zhang X, Lapata M. Sentence Simplification with Deep Reinforcement Learning. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. New Brunswick, NJ: Association for Computational Linguistics; 2017 Presented at: The 2017 Conference on Empirical Methods in Natural Language Processing; September 7-11, 2017; Copenhagen, Denmark p. 584-594 URL: <https://aclanthology.org/D17-1062.pdf> [doi: [10.18653/v1/d17-1062](https://doi.org/10.18653/v1/d17-1062)]
  12. Xu W, Napoles C, Pavlick E, Chen Q, Callison-Burch C. Optimizing Statistical Machine Translation for Text Simplification. TACL 2016 Dec;4:401-415. [doi: [10.1162/tacl\\_a\\_00107](https://doi.org/10.1162/tacl_a_00107)]
  13. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, et al. Attention is all you need. In: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc; 2017 Presented at: NIPS'17: The 31st International Conference on Neural Information Processing Systems; December 4-9, 2017; Long Beach, CA p. 6000-6010 URL: <https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>
  14. Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, et al. 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. New Brunswick, NJ: Association for Computational Linguistics; 2020 Jul Presented at: The 58th Annual Meeting of the Association for Computational Linguistics; July 5-10, 2020; Virtual p. 7871-7880 URL: <https://aclanthology.org/2020.acl-main.703.pdf> [doi: [10.18653/v1/2020.acl-main.703](https://doi.org/10.18653/v1/2020.acl-main.703)]
  15. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. Amazon AWS. 2022. URL: [https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language\\_understanding\\_paper.pdf](https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf) [accessed 2022-10-31]
  16. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Journal of Machine Learning Research 2020;21:1-67 [FREE Full text]
  17. Martin L, Fan A, de la Clergerie E, Bordes A, Sagot B. MUSS: multilingual unsupervised sentence simplification by mining paraphrases. arXiv Preprint posted online on April 16, 2021. [doi: [10.48550/arXiv.2005.00352](https://doi.org/10.48550/arXiv.2005.00352)]



18. Wenzek G, Lachaux MA, Conneau A, Chaudhary V, Guzmán F, Joulin A, et al. CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. In: Proceedings of the Twelfth Language Resources and Evaluation Conference.: European Language Resources Association; 2020 Presented at: LREC 2020: The 12th Conference on Language Resources and Evaluation; May 11-16, 2020; Marseille, France p. 4003-4012 URL: <https://aclanthology.org/2020.lrec-1.494>
19. Zhao Y, Chen L, Chen Z, Yu K. Semi-Supervised Text Simplification with Back-Translation and Asymmetric Denoising Autoencoders. AAAI 2020 Apr 03;34(05):9668-9675. [doi: [10.1609/aaai.v34i05.6515](https://doi.org/10.1609/aaai.v34i05.6515)]
20. Surya S, Mishra A, Laha A, Jain P, Sankaranarayanan K. Unsupervised Neural Text Simplification. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. New Brunswick, NJ: Association for Computational Linguistics; 2019 Presented at: The 57th Annual Meeting of the Association for Computational Linguistics; July 28-August 2, 2019; Florence, Italy p. 2058-2068 URL: <https://aclanthology.org/P19-1198.pdf> [doi: [10.18653/v1/p19-1198](https://doi.org/10.18653/v1/p19-1198)]
21. Sun R, Jin H, Wan X. Document-Level Text Simplification: Dataset, Criteria and Baseline. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. New Brunswick, NJ: Association for Computational Linguistics; 2021 Presented at: The 2021 Conference on Empirical Methods in Natural Language Processing; November 7–11, 2021; Online and Punta Cana, Dominican Republic p. 7997-8013 URL: <https://aclanthology.org/2021.emnlp-main.630.pdf> [doi: [10.18653/v1/2021.emnlp-main.630](https://doi.org/10.18653/v1/2021.emnlp-main.630)]
22. Coster W, Kauchak D. Simple English Wikipedia: a new text simplification task. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. New Brunswick, NJ: Association for Computational Linguistics; 2011 Presented at: The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies; June 19-24, 2011; Portland, OR p. 665-669 URL: <https://aclanthology.org/P11-2117.pdf>
23. Jiang C, Maddela M, Lan W, Zhong Y. Neural CRF Model for Sentence Alignment in Text Simplification. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. New Brunswick, NJ: Association for Computational Linguistics; 2020 Jul Presented at: The 58th Annual Meeting of the Association for Computational Linguistics; July 5-10, 2020; Virtual p. 7943-7960 URL: <https://aclanthology.org/2020.acl-main.709.pdf> [doi: [10.18653/v1/2020.acl-main.709](https://doi.org/10.18653/v1/2020.acl-main.709)]
24. Xu W, Callison-Burch C, Napoles C. Problems in Current Text Simplification Research: New Data Can Help. TACL 2015 Dec;3:283-297. [doi: [10.1162/tacl\\_a\\_00139](https://doi.org/10.1162/tacl_a_00139)]
25. Bjerva J, Bos J, van der Goot R, Nissim M. The Meaning Factory: Formal Semantics for Recognizing Textual Entailment and Determining Semantic Similarity. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). New Brunswick, NJ: Association for Computational Linguistics; 2014 Presented at: The 8th International Workshop on Semantic Evaluation (SemEval 2014); August 23-24, 2014; Dublin, Ireland p. 642-646 URL: <https://aclanthology.org/S14-2114.pdf> [doi: [10.3115/v1/s14-2114](https://doi.org/10.3115/v1/s14-2114)]
26. Laban P, Schnabel T, Bennett P, Hearst M. Keep It Simple: Unsupervised Simplification of Multi-Paragraph Text. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). New Brunswick, NJ: Association for Computational Linguistics; 2021 Presented at: The 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing; August 1–6, 2021; Online p. 6365-6378 URL: <https://aclanthology.org/2021.acl-long.498.pdf> [doi: [10.18653/v1/2021.acl-long.498](https://doi.org/10.18653/v1/2021.acl-long.498)]
27. van den Bercken L, Sips RJ, Lofi C. Evaluating neural text simplification in the medical domain. New York, NY: Association for Computing Machinery (ACM); 2019 May Presented at: WWW '19: The World Wide Web Conference; May 13-17, 2019; San Francisco CA p. 3286-3292 URL: <https://dl.acm.org/doi/10.1145/3308558.3313630> [doi: [10.1145/3308558.3313630](https://doi.org/10.1145/3308558.3313630)]
28. Dataset. Github. URL: <https://github.com/AshOlogn/Paragraph-level-Simplification-of-Medical-Texts> [accessed 2022-10-31]
29. Kincaid JP, Fishburne Jr RP, Rogers RL, Chissom BS. Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for navy enlisted personnel. Naval Technical Training Command Millington TN Research Branch. 1975 Feb 1. URL: <https://apps.dtic.mil/sti/citations/ADA006655> [accessed 2022-10-31]
30. Papineni K, Roukos S, Ward T, Zhu W. Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. New Brunswick, NJ: Association for Computational Linguistics; 2002 Presented at: The 40th Annual Meeting of the Association for Computational Linguistics; July 7-12, 2002; Philadelphia, PA p. 311-318 URL: <https://aclanthology.org/P02-1040.pdf> [doi: [10.3115/1073083.1073135](https://doi.org/10.3115/1073083.1073135)]
31. Chen Q, Peng Y, Lu Z. BioSentVec: creating sentence embeddings for biomedical texts. New York, NY: IEEE; 2019 Presented at: 2019 IEEE International Conference on Healthcare Informatics (ICHI); June 10-13, 2019; Xi'an, China p. 1-15 URL: <https://ieeexplore.ieee.org/document/8904728> [doi: [10.1109/ICHI.2019.8904728](https://doi.org/10.1109/ICHI.2019.8904728)]
32. Breland H. Word Frequency and Word Difficulty: A Comparison of Counts in Four Corpora. Psychol Sci 2016 May 06;7(2):96-99 [FREE Full text] [doi: [10.1111/j.1467-9280.1996.tb00336.x](https://doi.org/10.1111/j.1467-9280.1996.tb00336.x)]
33. Narayan S, Cohen SB, Lapata M. 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. New Brunswick, NJ: Association for Computational Linguistics; 2018 Presented at: The 2018 Conference on Empirical Methods in Natural Language Processing; October 31-November 4, 2018; Brussels, Belgium p. 1797-1807 URL: <https://aclanthology.org/D18-1206.pdf> [doi: [10.18653/v1/d18-1206](https://doi.org/10.18653/v1/d18-1206)]

34. Nallapati R, Zhou B, dos Santos C, Gu' lçehre C, Xiang B. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. In: Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning. New Brunswick, NJ: Association for Computational Linguistics; 2016 Aug Presented at: The 20th SIGNLL Conference on Computational Natural Language Learning; August 7-12, 2016; Berlin, Germany p. 280-290 URL: <https://aclanthology.org/K16-1028.pdf> [doi: [10.18653/v1/k16-1028](https://doi.org/10.18653/v1/k16-1028)]
35. Qi W, Yan Y, Gong Y, Liu D, Duan N, Chen J, et al. ProphetNet: Predicting Future N-gram for Sequence-to-Sequence Pretraining. In: Findings of the Association for Computational Linguistics, EMNLP 2020. New Brunswick, NJ: Association for Computational Linguistics; 2020 Presented at: EMNLP 2020; November 16-20, 2020; Online p. 2401-2410 URL: <https://aclanthology.org/2020.findings-emnlp.217.pdf>
36. Ranzato M, Chopra S, Auli M, Zaremba W. Sequence Level Training with Recurrent Neural Networks. arXiv Preprint posted online on May 6, 2016. [FREE Full text]
37. Aghajanyan A, Shrivastava A, Gupta A, Goyal N, Zettlemoyer L, Gupta S. Better Fine-Tuning by Reducing Representational Collapse. 2020 Apr Presented at: International Conference on Learning Representations (ICLR 2020); April 26–30, 2020; Virtual URL: [https://www.researchgate.net/publication/343547031\\_Better\\_Fine-Tuning\\_by\\_Reducing\\_Representational\\_Collapse](https://www.researchgate.net/publication/343547031_Better_Fine-Tuning_by_Reducing_Representational_Collapse)
38. Williams R. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Mach Learn 1992 May;8(3-4):229-256 [FREE Full text] [doi: [10.1007/BF00992696](https://doi.org/10.1007/BF00992696)]
39. Rennie SJ, Marcheret E, Mroueh Y, Ross J, Goel V. Self-Critical Sequence Training for Image Captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York, NY: IEEE; 2017 Jul Presented at: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); July 21-26, 2017; Honolulu, HI p. 7008-7024 URL: [https://openaccess.thecvf.com/content\\_cvpr\\_2017/papers/Rennie\\_Self-Critical\\_Sequence\\_Training\\_CVPR\\_2017\\_paper.pdf](https://openaccess.thecvf.com/content_cvpr_2017/papers/Rennie_Self-Critical_Sequence_Training_CVPR_2017_paper.pdf) [doi: [10.1186/isrctn12348322](https://doi.org/10.1186/isrctn12348322)]
40. Spasic I, Nenadic G. Clinical Text Data in Machine Learning: Systematic Review. JMIR Med Inform 2020 Mar 31;8(3):e17984 [FREE Full text] [doi: [10.2196/17984](https://doi.org/10.2196/17984)] [Medline: [32229465](https://pubmed.ncbi.nlm.nih.gov/32229465/)]
41. Martin L, De, Sagot B, Bordes A. Controllable Sentence Simplification. In: InProceedings of the 12th Language Resources and Evaluation Conference. 2020 May 11 Presented at: In Proceedings of the Twelfth Language Resources and Evaluation Conference; 2020-05-11; France p. 4689-4698 URL: <https://aclanthology.org/2020.lrec-1.577/>
42. Yan YY, Hu F, Chen J, Bhendawade N, Ye T, Gong Y, et al. FastSeq: Make Sequence Generation Faster. 2021 Aug 01 Presented at: InProceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations. Aug 2021; 2022-08-01; Thailand p. 218-226 URL: <https://aclanthology.org/2021.acl-demo.26/> [doi: [10.18653/v1/2021.acl-demo.26](https://doi.org/10.18653/v1/2021.acl-demo.26)]
43. Paulus R, Xiong C, Socher R. A Deep Reinforced Model for Abstractive Summarization. 2018 Presented at: International Conference on Learning Representations (ICLR 2018); April 30 to May 3, 2018; Vancouver, BC URL: [https://www.researchgate.net/publication/316875315\\_A\\_Deep\\_Reinforced\\_Model\\_for\\_Abstractive\\_Summarization](https://www.researchgate.net/publication/316875315_A_Deep_Reinforced_Model_for_Abstractive_Summarization)
44. Lin CY. ROUGE: A Package for Automatic Evaluation of Summarie. New Brunswick, NJ: Association for Computational Linguistics; 2004 Presented at: Text Summarization Branches Out; July 25 and 6, 2004; Barcelona, Spain p. 74-81 URL: <https://aclanthology.org/W04-1013.pdf>
45. Yuan W, Neubig G, Liu P. BARTScore: Evaluating Generated Text as Text Generation. 2021 May 21 Presented at: Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021; December 6-14, 2021; Virtual p. 27263-27277 URL: <https://proceedings.neurips.cc/paper/2021/hash/e4d2b6e6fdeca3e60e0f1a62fee3d9dd-Abstract.html>
46. Zhang J, Zhao Y, Saleh M, Liu P. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. 2020 Jul 13 Presented at: InInternational Conference on Machine Learning. 2020; 2020-07-13; Virtual URL: <http://proceedings.mlr.press/v119/zhang20ae>
47. Loshchilov I, Hutter F. Decoupled Weight Decay Regularization. 2018 Sep 27 Presented at: International Conference on Learning Representations; 2018; Vancouver, Canada.

## Abbreviations

- ARI:** Automated Readability Index
- BERT:** bidirectional encoder representations from transformers
- FKGL:** Flesch-Kincaid Grade Level
- GPT:** generative pretraining transformer
- MLE:** maximum likelihood estimation
- KIS:** Keep it Simple
- Lml:** maximum likelihood loss
- LS:** lexical simplification
- LSTM:** long short-term memory
- MUSS:** multilingual unsupervised sentence simplification

**PLS:** plain language summary

**RFlesch:** FKGL reward

**RL:** reinforcement learning

*Edited by T Hao; submitted 18.03.22; peer-reviewed by T Zhang, S Kim, H Suominen; comments to author 27.06.22; revised version received 08.08.22; accepted 12.10.22; published 18.11.22*

*Please cite as:*

*Phatak A, Savage DW, Ohle R, Smith J, Mago V*

*Medical Text Simplification Using Reinforcement Learning (TESLEA): Deep Learning–Based Text Simplification Approach*  
*JMIR Med Inform 2022;10(11):e38095*

*URL: <https://medinform.jmir.org/2022/11/e38095>*

*doi: [10.2196/38095](https://doi.org/10.2196/38095)*

*PMID:*

©Atharva Phatak, David W Savage, Robert Ohle, Jonathan Smith, Vijay Mago. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org>), 18.11.2022. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on <https://medinform.jmir.org/>, as well as this copyright and license information must be included.