Text Simplification via Adaptive Teaching

Seyed Ali Bahrainian[1,2]*, Jonathan Dou[1]*, Carsten Eickhoff[1,2]
[1]Brown University
[2]University of Tübingen
bahrainian@brown.edu, xiangzhi_dou@brown.edu , carsten.eickhoff@uni-tuebingen.de

Abstract

Text simplification is the process of rewriting a piece of text using simpler vocabulary and grammatical structure in order to make the text more accessible and understandable for a larger audience.

In this paper, we introduce a new text simplification model based on the notion of adaptive teaching using a teacher network and a text generation network. We name this new model Simplification via Adaptive Teaching (SAT). Our proposed model sets a new state-of-the-art performance in terms of standard simplification metrics such as SARI and D-SARI with a significant improvement over the previous state of the art on the D-Wikipedia dataset and the WikiDoc benchmark dataset. Moreover, we conduct a human evaluation in terms of text simplicity, correctness, and fluency to substantiate SAT's performance.

1 Introduction

Text simplification is an important area of research which aims at improving the understandability and accessibility of written texts for individuals with limited reading skills and education level (e.g., children (Kajiwara et al., 2013), dyslexic people (Rello et al., 2013), those suffering from autism (Barbu et al., 2015)), non-native speakers of a language, and generally people without expertise in a specific domain (e.g., in legal or medical documents). That is, highly technical texts with specialised vocabulary can be made more accessible to a wider audience by simplifying the text.

Text simplification (Al-Thanyyan and Azmi, 2021) is generally categorized as (1) Lexical Simplification, (2) Syntactic Simplification, (3) Statistical Machine Translation style and finally, the more recent (4) Neural Machine Translation style using Transformer models (Vaswani et al., 2017).

Lexical modification of a text, involves substituting complex vocabulary with simpler equivalents (Saggion, 2017). On the other hand, syntactic simplification is another form of simplification which focuses on using simpler sentence structure and grammar and shortening sentences (Saggion, 2017). With the emergence of Transformer models, the process of text simplification is typically formulated as an end-to-end process learning from data (Blinova et al., 2023). We also use the latter approach to text simplification. However, in the design of our proposed Simplification via Adaptive Teaching model (SAT), we do use a teacher network that explicitly learns necessary lexical modifications.

In this paper, we focus on developing a new text simplification method by explicitly modelling replacement of complex vocabulary by simpler terms in an adaptive learning setting. In order to do so, we consider the text simplification problem as a translation task such that complex text is translated to its simplified equivalent. Moreover, our proposed model utilizes two neural networks, (1) a Transformer-based sequence-to-sequence model which maps a complex input text to its simplified version. (2) A feed-forward network which maps complex vocabulary to simpler equivalents and through an integrated loss function, indicates to the Transformer model, how much vocabulary replacement it should perform to yield optimal results. We call our model *Simplification via Adaptive Teaching (SAT)*.

The main contributions of this paper are as follows:

- We present SAT, a new simplification model which sets a new state of the art in document simplification in terms of standard simplification metrics.
- We conduct an extensive experimentation to show the merit of this research work.

* Equal Contribution

Findings of the Association for Computational Linguistics: ACL 2024, pages 6574–6584

- We will release our code and pre-trained models in the final version of this paper in order to contribute to the reproducibility of our results and advance document simplification research.

The organization of the remainder of this paper is as follows: Section 2 discusses the related work briefly. In Section 3, we introduce our new model SAT along with its architecture. Subsequently, in Section 5, we show our model sets a new state of the art in terms of standard metrics by outperforming baseline models by considerable margins. Finally, we conclude this paper and present an outlook on future work before discussing pertinent limitations.

2 Related Work

Text simplification can be seen as a special form of controlled text generation (Bahrainian et al., 2022, 2021b) where a given input text is converted to its simplified equivalent. Recent text simplification research has been more focused on sentence simplification (Sheang and Saggion, 2021; Martin et al., 2021), which deal with simplification of short single sentence texts. The most commonly used datasets for text simplification such as WikiLarge (Zhang and Lapata, 2017), TurkCorpus (Xu et al., 2016a), and Newsela (Xu et al., 2015) are originally designed for sentence simplification. On the other hand, various real-world applications require document-level simplification rather than sentence-level processing. This stems from the necessity to understand the main ideas across multiple sentences simultaneously and rewrite them in a simplified vocabulary and grammar structure without respecting a number of sentences. Thus, document-level text simplification may have more applications than text simplification at the sentence level.

Sun et al. (2021) investigated the task of document-level text simplification, provided a large-scale dataset called D-Wikipedia, and proposed a more suitable evaluation metric than SARI (Xu et al., 2016b) named D-SARI in the document-level simplification task. Blinova et al. (2023) further preprocessed and cleaned the two datasets to remove faulty samples from both datasets. Furthermore, they also introduced a document-level simplification model named SimSum(Blinova et al., 2023) for document-level simplification. This model consists of a two stage setup of first summarizing and then simplifying a document end-to-end. In this paper, we also focus on document-level text simplification.

Analogous to most other NLP tasks, recent text simplification research has heavily relied on Transformer-based models (Vaswani et al., 2017). This recent research trend considers text simplification as a sequence-to-sequence problem, resembling machine translation (Narayan and Gardent, 2014) or document summarization (Liu et al., 2022). As a result, language models such as T5 (Raffel et al., 2020) or BART (Lewis et al., 2019) have been extensively used in text simplification research. For instance, the current state of the art on text simplification is Control_Prefixes(Clive et al., 2021), which is developed on an underlying T5 model as well as a BART-based variant. Due to the success of these three models we compare our new model SAT, against all three models. The Control_Prefixes model uses a dynamic method which allows for the inclusion of conditional input-dependent information, combining the benefits of prompt tuning and controlled generation. This method incorporates attribute-level learnable representations into different layers of a pre-trained Transformer, namely either T5 or BART. Another recent top-performing baseline is BRIO (Liu et al., 2022). This method holds the current state of the art for abstractive summarization, outperforming BART and and T5 by significant margins.

Due to the fact that abstractive summarization and document simplification share notable similarities, such as both being sequence-to-sequence NLP tasks, and both aiming at generating the main concepts and the gist of any given input article, we also compare our simplification model against this model. BRIO, unlike most abstractive summarization models which assume a deterministic (one-point) target distribution in which an ideal model will assign all the probability mass to the reference summary, proposes a novel training paradigm which assumes a non-deterministic distribution so that different candidate summaries are assigned probability mass according to their quality. The authors show that their method performs well during inference.

Another framework which has shown to achieve state-of-the-art performance on sentence-to-sentence simplification tasks is MUSS (Martin et al., 2021). MUSS utilizes sentence-level paraphrase data instead of simplification data for training. Furthermore, it leverages unsupervised pre-

training and controllable generation mechanisms to flexibly adjust attributes such as length and lexical complexity at inference time.

Other efforts focus on the correctness and factual accuracy in automatic text simplification by introducing a taxonomy of frequent errors (Devaraj et al., 2022).

In this paper, we propose a novel model using an adaptive teaching framework where a Transformer-based language model learns from a teacher network, the degree to which it needs to modify an input document in order to produce a simplified version of the input.

3 Proposed New Model

Our new model, SAT, consists of two neural networks: (1) a Transformer network based on an encoder-decoder architecture for sequence-to-sequence generation and (2) a teacher network which is a feed-forward neural network. The two networks are trained in an end-to-end fashion.

Figure 1 illustrates the SAT architecture. We initialize the Transformer network using a pre-trained BART model. The teacher network is initialized with random weights as a feed-forward network optimizing cross-entropy loss. Both networks receive as input the same documents written in potentially complex language (i.e. in terms of vocabulary and discourse), and as the target of the training they receive the corresponding documents in simplified language. Whether a text is complex or simple, is determined by human-annotated datasets explained in Section 5.1.

We elaborate on the details of each network and describe the learning process. The Transformer network uses an encoder-decoder architecture to map complex text to its simplified equivalent in a sequence-to-sequence text generation setup. On the other hand, the teacher network receives as input documents in potentially complex vocabulary in the form of a binary bag-of-words feature set, and the target of its training is mapping this input to its simplified corresponding document, again in the form of binary bag-of-words. Therefore, the feed-forward teacher network learns a mapping between complex vocabulary and its simplified equivalent via a cross-entropy loss function in a supervised fashion. After the teacher network is trained, it is integrated into the Transformer model to force the Transformer to use simple vocabulary. In order to do that, the loss function of the Transformer network is modified as follows. The Transformer model aims to learn a mapping between complex and simplified text that results in generating high-quality simplified text. In order to achieve this goal it uses a Maximum likelihood estimation (MLE) step. That is, it aims to maximize the likelihood of the reference simple text, S^*:

$$L^* = \underset{\theta}{\operatorname{argmax}} \sum_i log p_{g\theta}(S^{*(i)}|D^{(i)};\theta) \quad (1)$$

where L denotes the Transformer loss parametrized by g and $p_{g\theta}$ denotes the probability distribution entailed by these parameters. The summation is over the training set, and $\{D(i), S^{*(i)}\}$ is the i_{th} document and its simplified equivalent training sample.

The cross-entropy loss, $L_{teacher}$, of the teacher feed-forward network for $\{D^{(i)}, S^{*(i)}\}$ is computed, and a new loss term for the Transformer network is derived at inference time for each sample $\{D^{(i)}, S^{*(i)}\}$, such that:

$$\hat{L^*} = L^* + L_{teacher} \quad (2)$$

At training time, the teacher network is initially trained for a number of epochs. Subsequently, the Transformer network is trained using Equation 2. That means that each document in complex language is split into inputs and given to both the Transformer and teacher networks; first, the teacher network loss is computed and subsequently, the Transformer loss is updated. The teacher loss value indicates the degree of vocabulary simplification needed to reach the lowest loss, and therefore a reasonable simplified text. This loss value is directly added to the Transformer network's loss before error backpropagation. This is to encourage SAT to simplify complex words more than simpler ones.

At inference time, the Transformer model takes in the complex input documents and generates simplified versions. The teacher network is no longer used in this phase. The pairs of system outputs and ground truth simple texts are then used for evaluation.

3.1 Intuition Behind SAT

In this subsection we present further details about the model and explain its mechanics intuitively.

As discussed in the previous section, the feed-forward network takes as input a document representation in the form of a binary vector. Converting

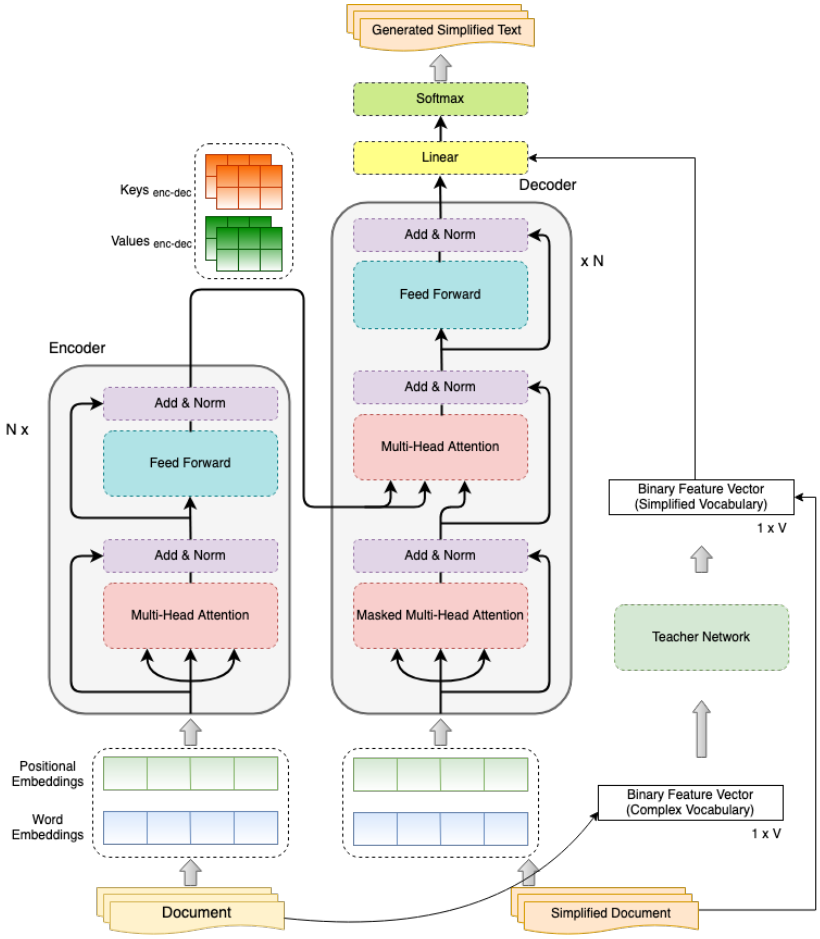

Figure 1: SAT model architecture, consisting of a Transformer network and a Teacher network. V denotes the vocabulary size. An input document is processed by both the Transformer network and the Teacher network. The two networks are then combined via a linear layer.

an input text to a binary vector involves: (1) computing and listing the entire vocabulary in a training dataset. (2) initializing a single dimensional vector in the size of the vocabulary with all entries being set to zero. This means that each word from the vocabulary has a corresponding value (at a specific index in the vector) which currently is set to 0. (3) going through the input document and for each word in the document, setting its corresponding value in the initialized vector to 1. The resulting vector is in the size of the vocabulary where the values associated with the words from the document are set to 1, while words that do not appear in the input document have a score of 0. The same process happens for the target simplified text during the training time.

Once an input document is converted to a binary vector, it is given to the teacher network. At the same time the document (i.e. without any modifications) is processed by the Transformer network. During the training, the teacher network computes a loss score by comparing the vector of an input text and its corresponding simplified target text. Then using Equation 2, the teacher loss is combined with the Transformer loss for the same document (or the same batch of documents).

During training, the teacher network outputs loss values proportional to the complexity of sentences, which is then directly added to the Transformer network loss to encourage further degrees of simplification by the Transformer network. A sentence such as "On this Tuesday afternoon, Bob is walking his German Shepherd on Fifth Avenue with a bright blue leash", would produce a higher loss value from the teacher network than the sentence "Bob is walking his dog today".

Therefore, it can be interpreted that the feed-forward network loss, plays the role of a teacher indicating to the Transformer network how much simplification is needed for each input document, hence the name SAT or Simplification via Adaptive Teaching. Previous work (Bražinskas et al.,

2020) has also utilized a similar training concept of using an auxiliary network to train the main text generation network more effectively. We believe that an extension of such training concept can have the potential of generating text satisfying various training objectives and criteria. A natural future work would be to study the integration of multiple teacher networks.

4 Detailed Explanation of Model Training and Testing

The SAT model is composed of a feed-forward teacher network embedded in a Transformer model. In this section we describe the model in detail and elaborate on the training and testing phases. We will release our implementation of SAT in the final version of this paper.

The teacher network contains one hidden linear layer followed by a softmax layer. Its purpose is to determine the complexities of sentences and map them to corresponding loss values, which are used to improve training and target more difficult simplification tasks. This is achieved by inputting binary encoded vectors representing the vocabulary in source (complex) sentences and expecting outputs of similar binary encoded vectors representing the vocabulary in target (simplified) sentences. We create the inputs and labels for the teacher network by converting source and simplified documents in the training data to binary encoded vectors of size 15869, matching the vocabulary size. Cross-entropy loss is calculated on the outputs and labels after softmax activation; higher loss values should indicate higher degrees of simplification needed. After pre-training the feed-forward network for 7 epochs, it is integrated into the Transformer model to assist training. As it can be seen in the model architecture in Figure 1, there is a linear layer combining the loss terms of both networks. That means that a forward-pass of both the Teacher network and the Transformer network does go through the linear layer, both loss terms are combined and then the calculated total loss is back-propagated through the linear layer onto the Transformer network, thus the gradients of the Transformer network are influenced by the forward pass of the Teacher network and updated accordingly. Finally, model parameters are determined using validation sets.

The Transformer model implements an encoder-decoder architecture initialized by BART. Its purpose is to process complex sentences and generate their simplified versions.

4.1 Training Details

At training time, we tokenize sentences and generate binary vectors to create inputs and targets for the teacher network. We feed each input document into both the Transformer (in the form of raw text) and the teacher network (in the form of binary vector), summing their losses before error backpropagation on the Transformer network. This encourages the Transformer model to target more complex sentences and simplify them to a further degree. The Transformer model is fine-tuned for 3 epochs. We set the number of epochs empirically, using the validation datasets.

4.2 Inference Details

At inference time, we input complex sentences from documents into the SAT model and use beam search to generate simplified sentences from the model's outputs. These simplified sentences are subsequently concatenated to form a simplified document. We note that the teacher network is not active in this phase. Instead, the Transformer model processes the input documents to produce their simplified counterparts. The beam size selected for our experiments is 4, determined to be optimal based on the results obtained.

5 Evaluation

We base our evaluation on common standard metrics (Sun et al., 2021) for text simplification, namely, SARI(Xu et al., 2016b), D-SARI(Sun et al., 2021), and BLEU(Papineni et al., 2002). We use EASSE (Alva-Manchego et al., 2019), a Python3 package created to standardize the evaluation of text simplification methods.

SARI (Xu et al., 2016b) compares the system output against references and the input document, explicitly measuring the quality of words that are added, deleted, and kept by the systems. SARI is the most popular used metric for text simplification task.

D-SARI (Sun et al., 2021) is a modified SARI score, adding different penalty factors based on the length. It is targeted towards document-level text simplification tasks.

BLEU (Papineni et al., 2002) was originally designed for Machine Translation tasks. It com-

putes the similarity between the system output and the reference simple document.

FKGL (Kincaid et al., 1975) is used to measure readability, but does not take grammar and semantic meanings into account.

5.1 Datasets

The most common dataset for evaluating document simplification is the D-Wikipedia dataset Sun et al. (2021). The second dataset that we use is the Wiki-doc collection (Kauchak, 2013) of Wikipedia articles. We use the versions of these datasets as released by `https://github.com/epfml/easy-summary` which are cleaned by removing misaligned and noisy samples (Blinova et al., 2023). The D-Wikipedia contains 97,074 training samples, 2,183 validation samples, and 5,836 test samples. Furthermore, Wiki-Doc contains 13,973 training samples, 1,768 validation samples, and 1,704 test samples.

5.2 Baselines

In this section, we list our baseline models, which also include the state of the art models for sentence-level simplification, document-level simplification, and document summarization:

BART (Lewis et al., 2019) is a top-performing pre-trained language model fine-tuned on a large corpus that has shown merit in various sequence-to-sequence tasks including the text simplification. Here we select the BART-base version.

T5 (Raffel et al., 2019) is an encoder-decoder language model pre-trained on a multi-task mixture of unsupervised and supervised tasks (each task is converted into a text-to-text format). Here we use the T5-base version.

BRIO (Liu et al., 2022) is another pre-trained model with outstanding performances for abstractive summarization. Here we select the model's checkpoint provided by the authors(`Yale-LILY/brio-cnndm-uncased`) and fine-tune it for our task.

MUSS (Martin et al., 2021) is a multilingual unsupervised sentence simplification system that trains strong models using sentence-level paraphrase data and achieves state-of-the-art results on sentence-level simplification task.

Control_Prefixes (Clive et al., 2021) uses a dynamic method which allows for the inclusion of conditional input-dependent information, and prompt tuning. It defines the state of the art on document-level simplification prior to SAT.

SimSum (Blinova et al., 2023) is a recent two-stage framework for automated document-level text simplification. This model is designed based on the notion of simplification via summarization. It incorporates explicit summarization and simplification models and guides the generation using the main keywords of a source text. It achieves very strong results on simplification tasks.

5.3 Experimental Results

In this section, we discuss the results of our experiments.

5.3.1 Automatic Evaluation

As mentioned previously, SAT is based on a BART-base model, and is fine-tuned in an adaptive teaching setting for 3 epochs as determined by the validation sets. After fine-tuning the hyperparameters for SAT, we found that a batch size of 16, decoder length of 250, and learning rate of 3^{-5} performed well on the validation sets. The teacher feed-forward network has a single layer followed by a softmax function. The inputs and outputs of this network are of the same size (i.e., the stop-word-filtered vocabulary size is $15,869$). The feed-forward network was trained for 7 epochs.

Table 1 presents the results of our experiments comparing SAT against top performing baselines in terms of SARI, D-SARI, BLEU, as well as, FKGL. Except for the unsupervised MUSS model, all models have been fully fine-tuned on the two datasets and were gone through due diligence in selecting the hyper-parameters using the validation sets. As we observe from the table, SAT significantly outperforms all the baselines in terms of SARI, D-SARI, and BLEU on both D-Wikipedia and Wiki-Doc datasets. In terms of FKGL, SAT outperforms all baseline models by achieving the lowest score on the Wiki-Doc dataset (i.e. Lower FKGL score is better). However, on the D-Wikipedia dataset we obseve that SAT outperforms all baseline models with the exception of SimSum.

In order to verify the statistical significance of our results, we conduct a significance testing. For each metric of SARI, D-SARI, and FKGL we conducted a paired test with SAT and in each case and on both datasets the best model after SAT. On the D-Wikipedia dataset, for SARI, D-SARI and FKGL this was the SIMSUM model. On the Wiki-Doc dataset for SARI and D-SARI this was the

	D-Wikipedia				Wiki-Doc			
model	SARI↑	D-SARI↑	BLEU↑	FKGL↓	SARI↑	D-SARI↑	BLEU↑	FKGL↓
T5	45.64	36.23	30.2	8.36	52.48	45.17	31.45	6.79
BART	47.05	38.13	24.21	8.14	53.71	46.3	30.99	7.93
BRIO	48.24	29.86	26.62	6.39	48.37	29.96	23.51	6.84
Control_Prefixes	49.16	38.29	27.81	7.12	54.67	46.39	32.41	7.08
MUSS	39.45	26.43	18.25	12.72	35.99	27.94	10.83	10.91
SIMSUM	49.44	39.77	28.41	**6.04**	49.11	41.53	30.45	6.79
$SAT(Ours)$	**50.66**	**41.27**	**31.38**	6.32	**58.38**	**48.78**	**36.64**	**6.67**

Table 1: A comparison of our model SAT against previous state-of-the-art baselines on two benchmark datasets. In the case of SARI, D-SARI, and BLEU higher numbers are better. In the case of FKGL lower numbers are better.

Control_Prefixes model and for FKGL this was SIMSUM. In order to overcome the issue with an unknown distribution we conducted the Mann-Whitney U test with a significance of 0.05 and p-value < 0.05. In the case of SARI and D-SARI, SAT's outputs' superiority on both datasets is statistically significant. In terms of FKGL, our comparison indicates that on D-Wikipedia the superiority of SIMSUM over SAT is statistically significant. However, on the Wiki-Doc dataset there is no statistically significant difference between the two model outputs, thus they are equal in FKGL performance. We conclude that SAT overall is the superior model as it sets a new state of the art on the main benchmark datasets for text simplification, being compared against prior state of the art with the exception of FKGL on the D-Wikipedia dataset.

We also made a comparison between SAT and Llama-2 7b and Llama-2 13b parameter models used for text simplification without any fine-tuning. The models were prompted with "Simplify this text:" followed by a test document. SAT outperformed both models in terms of all the simplification metrics used in this paper.

In a parallel study, we have deployed SAT as a document simplification tool in order to simplify informed consent forms in a medical domain. We have observed that SAT performs extremely well both in quantitative metrics such as shown in Table 1, but also produces high-quality simplified text that demonstrates its superior capabilities despite its small model size and simple architecture. In the following subsections we further conduct extensive evaluations to show the merit of this new model.

5.3.2 Human Evaluation

We also conducted a human evaluation on 100 randomly taken sample documents, 50 from each dataset comparing the outputs of SAT against T5, BART and Control_Prefixes in terms of simplicity, correctness, and fluency.

To elaborate on the details of the human study, we selected three aspects to define our evaluation criteria: (1) Simplicity (**S**): is the output simpler than the original document?, (2) Correctness (**C**): Does the output have factual errors compared to the original document?, and (3) Fluency (**F**): is the output grammatically correct and well-formed?

We asked two human evaluators to conduct this study. The two judges were native/proficient English speakers who were undergraduate students.

Our main motivation behind evaluating Simplicity, Correctness, and Fluency in this human study was: (1) We evaluated Simplicity as it was the most important metric for text simplification indicating the degree of simplification according to human perception. (2) We evaluated Correctness, since while text simplification models simplify text, occasionally they also distort the original information and modify it in ways that the semantics change. By evaluating correctness, we ensured that simplification does not come at the cost of compromising the correctness of the information according to the original document. (3) We also evaluated fluency, as it is an important factor in readability of a text and therefore fundamental to text simplification. The judges were undergraduate students who evaluated the texts.

Two human evaluators rate each output on a scale of 1 to 5 with 1 being the lowest. The evaluation scores of the two judges were then aggregated by averaging. Table 2 shows the results of this evaluation. The overall Cohen's Kappa agreement between the human evaluators was 0.63. The human evaluation indicates that SAT outperforms the baseline models in this qualitative assessment ac-

Model	S	C	F
T5	2.2	3.8	4.0
BART	2.6	**4.4**	4.5
Control_Prefixes	3.2	4.1	4.4
SIMSUM	3.9	4.2	**4.7**
SAT	**4.2**	**4.4**	**4.7**

Table 2: Human evaluation average results on D-Wikipedia and Wiki-Doc. S, C, and F denote Simplicity, Correctness, and Fluency, respectively.

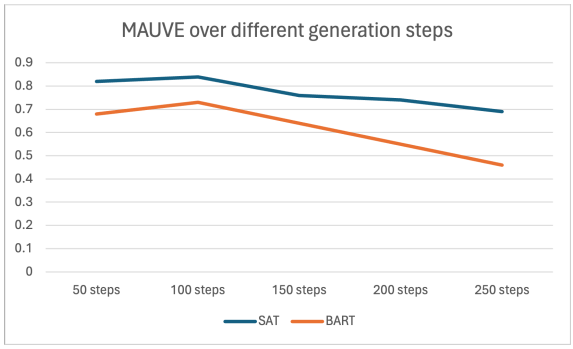

Figure 2: Ablation study comparing SAT against BART in terms of MAUVE with respect to different generation lengths

cording to human perception.

5.3.3 Ablation Studies on the Teacher Network

Effect of the Teacher Network on Simplification Metrics: The first ablation study that we conduct is to compare SAT against BART. The results of this comparison are presented in Table 1. As previously explained, SAT is initialized using a BART model, therefore, a head-to-head comparison between these two models indicates that the teacher network in SAT does provide a valuable addition to the models performance. That can be observed on all evaluation metrics, namely, SARI, D-SARI, BLEU, and FKGL.

Effect of the Teacher Network in terms of MAUVE: As a second ablation study we compare SAT against SAT without the teacher network (i.e. that means a BART model fine-tuned on the D-Wikipedia dataset) in terms of MAUVE (Pillutla et al., 2021). Mauve is an evaluation metric which by computing information divergences in a quantized embedding space, it directly compares the learnt distribution from a text generation model to the distribution of human-written groundtruth. In order to conduct this experiment, we use the validation set of the D-Wikipedia dataset and generate its samples at different generation lengths (i.e. steps). We test the generation lengths at cut-offs of 50, 100, 150, 200, and 250. The remaining parameters are kept the same as reported in Section 5.3.1. Figure 2 presents the results of this comparison. It can be observed SAT consistently outperforms BART indicating that is a quantized embedding space, SAT's outputs are closer to that of groundtruth as compared with BART. More notably, SAT stays more consistent as the generation length increases while the MAUVE score for BART drops more rapidly.

Summary of Other Ablation Studies on the Teacher Network: We conducted multiple ablations on the validation set of the D-Wikipedia dataset and found out that the tested settings did not improve SAT's overall performance in terms of the simplification metrics. Therefore, we just summarize these experiments bellow: (1) We modified the teacher network from the current single-layer to a 2-layer feed forward network of the same size with RELU activation functions, however, this did decrease performance. (2) We used RELU activations in a single-layer network instead of the current linear activations which also dropped performance. (3) Finally, we lemmatized the words inputted to the teacher network which also resulted in a poorer performance.

Based on the results of these two ablation studies we conclude that the notion of adaptive teaching is one that can be beneficial to language models in a sequence-to-sequence task.

5.3.4 Qualitative Analysis of Sample Outputs

Table 3 shows two qualitative examples of SAT output. In the examples, we observe a few edit operations such as word removal, word replacement (highlighted) and sentence shortening. We observe that for the input document "This category contains the talk page of articles which relate to living persons.", SAT replaces the phrase "the talk page of articles" with its simplified equivalent which is "articles". We also note that BART fails to maintain context and semantics by changing "the talk page of articles which relate to living persons" to "people who are living in the United States" while SAT preserves the ground truth with its simplification. We observe that even a large foundation model such as GPT-4o has difficulty avoiding complex vocabulary, demonstrated by its use of the word "continuously" which is more complex than "over and over". Finally, while SIMSUM gener-

Complex Texts	This category contains the talk page of articles which relate to living persons.	To roll means to move along a surface by revolving over and over.
BART	This category contains people who are living in the United States.	To roll means to move along a surface by turning over and over again.
SIMSUM	This category has the talk pages for articles relating to living people.	To roll means to move along a surface over and over again.
GPT-4o	This category contains the discussion pages of articles about living people.	To roll means to move by continuously turning over on a surface.
SAT	This category contains articles which relate to living persons.	To roll means to move along a surface by turning over and over.

Table 3: Complex text and their simplified equivalent model-generated outputs comparing SAT against various models. Modifications are highlighted.

ates simple outputs, it fails to preserve semantics in the second example by its removal of "revolving" without a proper substitution.

6 Conclusions and Future Work

In this paper, we proposed SAT, an innovative adaptive teaching approach to document simplification. SAT outperforms existing top baselines by a considerable margin in simplification scores. The architecture of SAT is composed of a feed-forward Teacher network integrated into a Transformer sequence-to-sequence network via a joint loss function. Our model has shown merit as a practical text simplification tool, not only in the results demonstrated in this paper but also in parallel applied research work simplifying informed consent forms in the healthcare domain. We showed through extensive experimentation both on automatic evaluation metrics such as SARI, D-SARI, BLEU, and FKGL, as well as, human evaluations of simplicity, correctness, and fluency that this light-weight model has a remarkable performance achieving top results. Moreover, we conducted an extensive ablation study testing the effect of the teacher network and multiple different settings to corroborate the effectiveness of SAT model's design in its current form.

In the future, we plan to study controlled document simplification, tailored towards specific needs and preferences of various audiences, e.g., children, non-native speakers, dyslexic, etc. Another interesting direction for future exploration would be studying adding multiple teacher networks to build a more complex objective function for training the text generation network. Such objective function could for instance control the style or the topic-focus(Bahrainian et al., 2021a) of the final text output. This could open a path to studying such effects along-side various fine-tuning schemes beyond linear addition of loss terms.

7 Limitations

The limitations of our work include (1) a lack of studying different model sizes, although we demonstrate that with a base-size model, SAT can outperform large variants of models such as BRIO and MUSS. (2) The performance of Transformer-based simplification models including ours is dataset-dependant and does require cleaned datasets dedicated for simplification. (3) Document simplification tailored to a target audience is not studied in this work and we offer a single solution for all demographics in need for simplifying text for better understandability.

References

Suha S. Al-Thanyyan and Aqil M. Azmi. 2021. Automated text simplification: A survey. *ACM Comput. Surv.*, 54(2).

Fernando Alva-Manchego, Louis Martin, Carolina Scarton, and Lucia Specia. 2019. EASSE: Easier automatic sentence simplification evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 49–54, Hong Kong, China. Association for Computational Linguistics.

Seyed Ali Bahrainian, Sheridan Feucht, and Carsten Eickhoff. 2022. NEWTS: A corpus for news topic-focused summarization. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 493–503, Dublin, Ireland.

Seyed Ali Bahrainian, Martin Jaggi, and Carsten Eickhoff. 2021a. Self-supervised neural topic modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3341–3350, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Seyed Ali Bahrainian, George Zerveas, Fabio Crestani, and Carsten Eickhoff. 2021b. Cats: Customizable abstractive topic-based summarization. *ACM Trans. Inf. Syst.*, 40(1).

Eduard Barbu, M Teresa Martín-Valdivia, Eugenio Martinez-Camara, and L Alfonso Urena-López. 2015. Language technologies applied to document simplification for helping autistic people. *Expert Systems with Applications*, 42(12):5076–5086.

Sofia Blinova, Xinyu Zhou, Martin Jaggi, Carsten Eickhoff, and Seyed Ali Bahrainian. 2023. SIMSUM: Document-level text simplification via simultaneous summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9927–9944, Toronto, Canada. Association for Computational Linguistics.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020. Few-shot learning for opinion summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4119–4135, Online. Association for Computational Linguistics.

Jordan Clive, Kris Cao, and Marek Rei. 2021. Control prefixes for parameter-efficient text generation.

Ashwin Devaraj, William Sheffield, Byron Wallace, and Junyi Jessy Li. 2022. Evaluating factuality in text simplification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7331–7345, Dublin, Ireland. Association for Computational Linguistics.

Tomoyuki Kajiwara, Hiroshi Matsumoto, and Kazuhide Yamamoto. 2013. Selecting proper lexical paraphrase for children. In *Proceedings of the 25th Conference on Computational Linguistics and Speech Processing (ROCLING 2013)*, pages 59–73, Kaohsiung, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

David Kauchak. 2013. Improving text simplification language modeling using unsimplified text data. In *Proceedings of the 51st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 1537–1546.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. Brio: Bringing order to abstractive summarization.

Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2021. Muss: Multilingual unsupervised sentence simplification by mining paraphrases. *arXiv preprint arXiv:2005.00352*.

Shashi Narayan and Claire Gardent. 2014. Hybrid simplification using deep semantics and machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 435–445, Baltimore, Maryland. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*, 34:4816–4828.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer.

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. *Journal of Machine Learning Research*, 21(140):1–67.

Luz Rello, Ricardo Baeza-Yates, Stefan Bott, and Horacio Saggion. 2013. Simplify or help? text simplification strategies for people with dyslexia. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, pages 1–10.

Horacio Saggion. 2017. Automatic text simplification. *Synthesis Lectures on Human Language Technologies*, 10(1):1–137.

Kim Cheng Sheang and Horacio Saggion. 2021. Controllable sentence simplification with a unified text-to-text transfer transformer. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 341–352, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Renliang Sun, Hanqi Jin, and Xiaojun Wan. 2021. Document-level text simplification: Dataset, criteria and baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7997–8013, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016a. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016b. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594, Copenhagen, Denmark. Association for Computational Linguistics.