# SIERA: An Evaluation Metric for Text Simplification using the Ranking Model and Data Augmentation by Edit Operations

**Hikaru Yamanaka and Takenobu Tokunaga**
School of Computing, Tokyo Institute of Technology
Tokyo Meguro Ôokayama 2-12-1, Japan
hyamanak@lycorp.co.jp, take@c.titech.ac.jp

**Abstract**

Automatic evaluation metrics are indispensable for text simplification (TS) research. The past TS research adopts three evaluation aspects: fluency, meaning preservation and simplicity. However, there is little consensus on a metric to measure simplicity, a unique aspect of TS compared with other text generation tasks. In addition, many of the existing metrics require reference simplified texts for evaluation. Thus, the cost of collecting reference texts is also an issue. This study proposes a new automatic evaluation metric, SIERA, for sentence simplification. SIERA employs a ranking model for the order relation of simplicity, which is trained by pairs of the original and simplified sentences. It does not require reference sentences for either training or evaluation. The sentence pairs for training are further augmented by the proposed method that utilizes edit operations to generate intermediate sentences with the simplicity between the original and simplified sentences. Using three evaluation datasets for text simplification, we compare SIERA with other metrics by calculating the correlations between metric values and human ratings. The results showed SIERA's superiority over other metrics with a reservation that the quality of evaluation sentences is consistent with that of the training data.

## 1. Introduction

Text simplification (TS) rewrites texts into simple and understandable ones while retaining their original meaning (Alva-Manchego et al., 2021). TS is expected to be an assistive technology for readers like children, non-native speakers and people with reading difficulties (Gooding, 2022). Recent TS models can generate fluent sentences by leveraging neural machine translation techniques, transforming a complicated sentence to its simplified counterpart within the same language (Al-Thanyyan and Azmi, 2021).

The performance of TS systems has been evaluated in terms of the following three aspects (Martin et al., 2018; Alva-Manchego et al., 2020, 2021).

- Fluency: Is the simplified text natural and free from grammatical errors?
- Meaning preservation: Does the simplified text retain the core meaning of the original?
- Simplicity: Is the simplified text easier to understand than the original?

Fluency and meaning preservation are common evaluation aspects in text generation tasks in general, and several automatic evaluation metrics have been proposed (Sai et al., 2022). In particular, BLEU (Papineni et al., 2002) and BERTScore (Zhang et al., 2020) are popular metrics for evaluating fluency and meaning preservation of texts. In contrast, simplicity is a unique evaluation aspect of TS and indispensable for TS research.

Automatic evaluation metrics are classified into two categories: reference-based and reference-free metrics. Reference-based metrics utilize reference texts for calculating evaluation scores for the target text, while reference-free metrics do not. Evaluation metrics for the text generation tasks are often reference-based. However, collecting manually written references for evaluation is expensive and time-consuming. In addition, it is inappropriate to regard the reference texts as the only correct output since there can be multiple acceptable simplified texts. Against this backdrop, we develop a reference-free metric for simplicity in this study.

To evaluate simplicity in TS, several automatic evaluation metrics have been proposed, including both reference-free (Kincaid et al., 1975; Sulem et al., 2018b) and reference-based (Papineni et al., 2002; Xu et al., 2016; Zhang et al., 2020) methods. However, it has been reported that these existing metrics are inappropriate for evaluating simplicity because of low correlation with manual evaluation (Alva-Manchego et al., 2020, 2021; Scialom et al., 2021). The evaluation metric of simplicity in TS research is still an open problem.

In this study, we limit the scope of TS to a sentence and propose a novel reference-free metric for evaluation of sentence simplicity, which we call **SIERA** (SImplification metric based on Edit operation through learning to RAnk). SIERA requires only parallel corpora of original and simplified sentences for training the evaluation model. The references are not necessary for calculating the evaluation scores. Following the framework of previous reference-free trainable evaluation met-



*3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI)*, pages 47–58
20-25 May, 2024May 20th, 2024. 

rics for other than text simplification (Wu et al., 2020; Maeda et al., 2022), the training procedure of SIERA consists of two parts: (1) learning-to-rank for determining the order relation of simplicity in training parallel corpora and (2) data augmentation to increase the number of training sentence pairs using edit operations between the original and simplified sentence pairs.
We summarize our contribution as follows.

- We propose a novel reference-free automatic evaluation metrics SIERA for sentence simplification, which can be trained only by a parallel corpus of original and simplified sentences.
- We develop a data augmentation method for extending the parallel corpus by considering edit operations between the original and simplified sentences.
- We demonstrate the superiority of SIERA to other automatic evaluation metrics for TS on three different evaluation datasets.

## 2. Related Work

**Reference-based metrics**
Reference-based metrics need reference sentences written by humans to evaluate simplified sentences. SARI (Xu et al., 2016), BLEU (Papineni et al., 2002) and BERTScore (Zhang et al., 2020) are common metrics in TS. SARI is a widely used metric for evaluating simplicity, which calculates the percentage of correctly added, kept, and deleted $n$-grams among the input, output and reference sentences. However, because SARI was initially proposed to evaluate lexical simplification, it is less suitable for evaluating simplified sentences with multiple rewriting operations (Alva-Manchego et al., 2020).
BLEU calculates a similarity score based on $n$-gram matching between output and reference sentences. Despite its simple computation and interpretability, BLEU is not recommended as a simplicity metric for sentences with splitting operations (Sulem et al., 2018a).
In contrast to BLEU and SARI, which rely on surface features like $n$-grams, BERTScore utilizes the BERT (Devlin et al., 2019) embeddings to compute sentence similarities considering contextual meaning. BERTScore aligns tokens of the output and reference sentences to calculate the cosine similarity between the aligned token embeddings. Alva-Manchego et al. (2021) reported that BERTScore is superior to BLEU and SARI in simplicity evaluation.
Recently, learnable reference-based automatic evaluation metrics have also been proposed. Maddela et al. (2023) developed LENS, which employs an adaptive ranking loss to weight reference sentences based on their similarity to the sentence to evaluate in terms of edit operation. The LENS metric correlates more with human evaluation than the previous reference-based metrics, such as SARI and BERTScore.
However, the reference-based methods have a drawback; collecting manually written references is expensive and time-consuming. Also, Alva-Manchego et al. (2021) pointed out that a high similarity to the reference does not necessarily indicate high simplicity since there can be acceptable sentences other than references, and manually written references have diverse levels of simplicity against the original sentences.

**Reference-free metrics**
Reference-free metrics evaluate sentences without references. SAMSA (Sulem et al., 2018b) calculates whether the semantic structure between the input and output sentences is maintained after sentence splitting. However, SAMSA focuses on simplification by sentence splitting; its evaluation performance is poor for simplification with multiple rewriting operations (Alva-Manchego et al., 2021).
FKGL (Kincaid et al., 1975) is another reference-free metric, which calculates from the average number of words and syllables. FKGL was initially proposed as a readability metric for grade levels in the United States, but it is often used to evaluate simplicity in TS. Tanprasert and Kauchak (2021) showed that FKGL is less robust against superficial edit operations, claiming that it is inappropriate for simplicity evaluation.
Vajjala and Meurers (2016) proposed the first pairwise ranking model to predict the readability of sentences. They created a classical ranking model that takes into account lexical and syntactic features to predict the readability of sentences and proposed to use it as an automatic evaluation metric of TS. Lee and Vajjala (2022) proposed a Neural Pairwise Ranking Model (NPRM) to predict sentence readability, which is a pairwise ranking model based on neural dense layers and BERT embedding. NPRM has not yet been investigated to see if it can be used for simplicity evaluation. We propose SIERA by extending the NPRM architecture.
More recently, Cripwell et al. (2023) proposed a learnable reference-free metric Simplicity Level Estimate (SLE) that calculates the absolute simplicity score of a single sentence. The goodness of simplification from the original to simplified sentences is calculated by the difference in their SLE scores. Unlike SLE, SIERA directly evaluates simplification for a given pair of original and simplified sentences.

**Edit operations**
Edit operations are often utilized in the simplification models. Alva-Manchego et al. (2017) pro-

posed a sequence transformation model that predicts edit operation tags such as deletion, replacement, and addition. Dong et al. (2019) extended this model to EditNTS, which performs edit operation prediction and adaptation in parallel, leveraging data that is automatically assigned token-by-token edit operations using dynamic programming. In recent years, a sentence simplification model has been proposed, which incrementally adds edit operations to improve a simplicity metric through unsupervised learning (Dehghan et al., 2022).

There is also a trend toward developing a typology of edit operations. Cardon et al. (2022) manually annotated a TS corpus with edit operations according to their thoughtful typology, suggesting the importance of TS evaluation in terms of fine-grained edit operation units. Yamaguchi et al. (2023) proposed a taxonomy of edit operations at the surface and content levels for understanding TS systems. Heineman et al. (2023) also organized 21 categories of edit operations for TS evaluation. Recently, there have been attempts to automatically generate these typologies of edit operations using LLM (Cardon and Bibal, 2023).

## 3. Resources

### 3.1. Training data

We use Newsela (Xu et al., 2015) as the training data for SIERA. Newsela is built upon data from 1,130 English news articles manually rewritten by professional editors in four levels of plain language to match the grade level of children, i.e. in principle, the original article has four variants corresponding to each simplicity level (1– 4), with 4 being the most simple level. The Newsela data is composed of parallel data aligned by sentences using Jaccard similarity. The number of total sentence pairs is 141,582.

### 3.2. Evaluation data

The evaluation data set for TS evaluation metrics consists of pairs of original and simplified text and manually assigned evaluation ratings for each pair regarding fluency, meaning preservation, and simplicity. The evaluation metrics are evaluated by measuring the correlation between these manual ratings and the evaluation scores obtained from the evaluation metrics. This study uses three sets of evaluation data, which are English corpora.

**Simplicity-DA**
Simplicity-DA (Alva-Manchego et al., 2021) is a data set consisting of original sentences from TurkCorpus (Xu et al., 2016) and corresponding simplified sentences created by six automatic simplification models[1]. Each model generated 100 simplified sentences, resulting in 600 sentence pairs. The manual ratings are assigned as continuous values ranging from 0 to 100. Fifteen ratings are collected for each sentence pair.

**Human-Likert**
Human-Likert (Scialom et al., 2021) also uses TurkCorpus as the original sentences, but unlike Simplicity-DA, the simplified sentences are written manually, comprising 100 sentence pairs in total. Thirty human ratings ranging from 0 to 100 are collected for each sentence pair.

**SimpDA$_{2022}$**
SimpDA$_{2022}$ (Maddela et al., 2023) uses source texts extracted from Wikipedia from 22/10/2022 to 24/11/2022, to evaluate long and complex sentences. These source sentences are simplified by two humans and four recent TS models[2], resulting in a total of 360 sentence pairs. Three manual ratings ranging from 0 to 100 were assigned to each sentence pair.

## 4. SIERA ranking model

We propose SIERA by extending NPRM (Lee and Vajjala, 2022). This section describes the outline of NPRM and its possible improvement. Then, we propose a SIERA ranking model based on the NPRM architecture.

### 4.1. NPRM

**Training** NPRM uses only parallel data consisting of original sentences and their simplified sentences during training. Let $n$ be the total number of the original sentences, $s_i$ be the $i$-th original sentence, and $s'_i$ be the corresponding simplified sentence. The instances for training data are made by concatenating $s_i$ and $s'_i$ by separating a SEP token in both orders as in (1). The arrow over $p_i$ indicates the order of the original and simplified sentences in the pair, i.e. the source of the arrow indicates the original sentence.

$$\begin{aligned} \overrightarrow{p_i} &= \textit{concat}(s_i; \text{SEP}; s'_i) \\ \overleftarrow{p_i} &= \textit{concat}(s'_i; \text{SEP}; s_i) \end{aligned} \tag{1}$$

We use notation $p_i$ for denoting both $\overrightarrow{p_i}$ and $\overleftarrow{p_i}$. The expected correct label $y_i$ for $p_i$ is either row vector $[0, 1]$ for $\overrightarrow{p_i}$ or $[1, 0]$ for $\overleftarrow{p_i}$. The element value $1$ indicates the simplified sentence position in a pair.

[1] ACCESS (Martin et al., 2020), DMASS-DCSS(Zhao et al., 2018), Dress-Ls (Zhang and Lapata, 2017), Hybrid (Narayan and Gardent, 2014), PBMT-R (Wubben et al., 2012) and SBMT-SARI (Xu et al., 2016).

[2] GPT-3.5 (Ouyang et al., 2022) w/ zero and few-shot, Muss (Martin et al., 2022) and T5-3B (Raffel et al., 2019).

NPRM calculate the output $o_i$ through BERT (Devlin et al., 2019) and a fully-connected feed-forward neural network (FFNN) as in (2), where BERT$(\cdot)$ denotes an output vector corresponding to the CLS token of BERT[3]. The output $o_i$ is a two-dimensional column vector, where each vector element represents the probability that the sentence corresponding to that element is a simplified setnence.

$$o_i = \textit{softmax}(\text{FFNN}(\text{BERT}(p_i))) \quad (2)$$

The created training data $p_i$ and corresponding labels $y_i$ are fed into the model and trained with a loss function (3),

$$L = -\sum_{i=1}^{n} y_i \cdot \log(o_i), \quad (3)$$

where $y_i$ and $o_i$ represents both $\overrightarrow{y_i}$ and $\overleftarrow{y_i}$, and both $\overrightarrow{o_i}$ and $\overleftarrow{o_i}$ respectively and correspondingly. $\log(o_i)$ denotes the element-wise application of the logarithmic function.

**Inference** Given an original sentence $s$ and its simplified sentence $s'$, NPRM calculates a readability score as in (6).

$$\overrightarrow{p} = \textit{concat}(s; \text{SEP}; s'), \quad (4)$$
$$\overrightarrow{o} = \textit{softmax}(\text{FFNN}(\text{BERT}(\overrightarrow{p}))), \quad (5)$$
$$\textit{readability score} = [0, 1] \cdot \overrightarrow{o}. \quad (6)$$

We can utlize this readability score to measure the simplicity of $s'$ against $s$.

### 4.2. Improvement of NPRM

Comparing the training phase ((1) and (2)) and the inference phase ((4), (5) and (6)) of NPRM, we find that NPRM utilizes both forwardly and backwardly-ordered pairs ($\overrightarrow{p_i}$ and $\overleftarrow{p_i}$) for training, but utilizes only the forwardly-ordered pairs for inference. We suspect that the inference phase of NPRM does not fully utilize the learned result.

We propose to utilize the backwardly-ordered sentence pair ($\overleftarrow{p}$) in addition to the forwardly-ordered sentence pair ($\overrightarrow{p}$) also in the inference phase to calculate the score as in (9). Equation (7) and (8) are the counterpart of (4) and (5), respectively.

$$\overleftarrow{p} = \textit{concat}(s'; \text{SEP}; s), \quad (7)$$
$$\overleftarrow{o} = \textit{softmax}(\text{FFNN}(\text{BERT}(\overleftarrow{p}))), \quad (8)$$
$$\textit{simplicity score} = \frac{1}{2}([0, 1] \cdot \overrightarrow{o} + [1, 0] \cdot \overleftarrow{o}). \quad (9)$$

[3]Although NPRM has freedom in the choice of neural network architectures; we adopt BERT and FFNN following Lee and Vajjala (2022)'s experimental setting.

## 5. Data Augmentation

This section describes a method to extend the parallel data for training the SIERA ranking model. We utilize edit operations for simplification to increase the original and simplified sentence pairs. Given a pair of an original sentence $s$ and its simplified sentence $s'$, the simplification can be represented by a set of edit operations that transform $s$ to $s'$. Alva-Manchego et al. (2020) reported that applying more edit operations for simplification makes the resultant sentence simpler. Following their finding, we apply subsets of the edit operations that bridge between $s$ and $s'$ to create new sentences which are simpler than $s$ but less simple than $s'$. We call them *intermediate sentences*. Suppose we create an intermediate sentence $\hat{s}$ from $s$ by applying a subset of edit operations; we can create new sentence pairs ($s$, $\hat{s}$) and ($\hat{s}$, $s'$). Theoretically, if we can transform $s$ to $s'$ through $N$ operations, we could create $2^N - 2$ intermediate sentences; thus we obtain $2(2^N - 2)$ new sentence pairs for training the SIERA ranking model.

Following Dong et al. (2019), we consider two levels of the edit operation: *token unit edit operation* (TE) and *span unit edit operation* (SE). TE is an edit operation applied to each token in the original sentence to transform it into a simplified sentence. There are three types of TE: ADD token, DELete token, and KEEP token. To extract TEs from given sentence pairs, we adopt the implementation by Dong et al. (2019)[4]. SEs are constructed by concatenating consecutive TEs except for KEEP. There are following three types of SEs. Figure 1 shows an example of extracted TEs and SEs.

- ADD-DEL span: A span in which one or more consecutive ADDs and DELs are combined in this order. This corresponds to lexical simplification and sentence splitting.
- DEL span: One or more consecutive DEL spans other than the ADD-DEL span. This corresponds to the deletion of unnecessary information.
- ADD span: One or more consecutive ADD spans other than the ADD-DEL span. This corresponds to the addition of necessary information.

We create intermediate sentences by applying a subset of the extracted SEs to the original sentence. However, an arbitrary subset of the extracted SEs does not always produce a valid sentence. For instance, among four SEs in Figure 1,

[4]https://github.com/YueDongCS/EditNTS/blob/master/label_edits.py

| | |
|---|---|
| Original | According to Ledford , Northrop executives said they would build substantial parts of the bomber in Palmdale , creating about 1,500 jobs . |
| TEs | KEEP KEEP KEEP KEEP KEEP DEL KEEP KEEP KEEP KEEP ADD(most) DEL DEL KEEP KEEP KEEP ADD(parts) KEEP KEEP ADD(.) ADD(It) ADD(would) ADD(create) DEL DEL DEL KEEP KEEP KEEP |
| SEs | KEEP KEEP KEEP KEEP KEEP DEL KEEP KEEP KEEP KEEP ADD(most) DEL DEL KEEP KEEP KEEP ADD(parts) KEEP KEEP ADD(.) ADD(It) ADD(would) ADD(create) DEL DEL DEL KEEP KEEP KEEP |
| Simplified | According to Ledford, Northrop said they would build most of the bomber parts in Palmdale. It would create 1,500 jobs . |

Figure 1: Example of extracted TEs and SEs. Each highlighted color represents ADD-DEL span, DEL span, and ADD span. In this example, four SEs are extracted in total.

we can apply the first DEL SE and the last ADD-DEL SE independently, but the second and third SEs must be applied simultaneously to rewrite "substantial parts of the bomber" to "most of the bomber parts". Applying only one of them produces invalid sentences. Although we can theoretically create $2^4 - 2 = 14$ intermediate sentences from this example, invalid sentences should be excluded from them.

To exclude irrelevant intermediate sentences, we discard intermediate sentences dissimilar from the original and simplified sentences in terms of BERTScore (Zhang et al., 2020). More concretely, we calculate BERTscore$_{f1}$ of an intermediate sentence with its original and simplified sentence each and average them. These average scores are further averaged across the entire generated intermediate sentences to determine a threshold. We discard the intermediate sentences that have lower average scores than the threshold. We randomly choose $m$ sentences from the remaining intermediate sentences to augment the training sentence pairs.

## 6. Experiment

### 6.1. Experimental settings

**Training data and models to compare**

We use the Newsela dataset for training the SIERA ranking model. Newsela comprises original news articles and corresponding simplified variants over four simplification levels. We first train a baseline model (Base) using 16,084 sentence pairs of the original and its most simplified sentence in Newsela. Next, we extend the sentence pairs for the baseline model by our proposed augmentation method described in section 5, resulting in 38,120 sentence pairs in total. We adopt a single intermediate sentence for each original sentence pair, i.e. the hyperparameter $m = 1$. Theoretically, we should obtain three times the original number of sentence pairs, but we have fewer sentence pairs in reality due to the filtering process to exclude irrelevant intermediate sentences. We call the SIERA model trained with this extended data +Silver. Furthermore, we extend the sentence pairs for the baseline using manually simplified sentences of the intermediate level of Newsela, resulting in 46,470 sentence pairs. The difference from the +Silver's training data is that the quality of intermediate sentences is guaranteed because they are written by professional editors. Therefore, we do not apply filtering in this data augmentation. Despite no filtering, the total number of sentence pairs is slightly fewer than three times that of the Base training data. This is because some Newsela articles do not have simplified sentences of intermediate levels. We call the SIERA model trained with this extended data +Gold.

We also consider the variants of these three models in the inference phase. The SIERA model uses both forwardly and backwardly-ordered sentence pairs in the inference phase (a two-way model). We consider the models that use only one of them in the inference phase and denote them by putting an arrow over the model name, i.e. $\rightarrow$ and $\leftarrow$ indicate using only forwardly or backwardly-ordered sentence pairs, respectively (a one-way model). Note that $\overrightarrow{\text{Base}}$ is equivalent to NPRM.

We also consider the existing reference-based (SARI, BLEU[5] , BERTScore[6]) and reference-free metrics (SAMSA, FKGL)[7].

**Hyperparameters**

We used the bert-base-uncased[8] model from Huggingface Transformers as a pre-training model and a ranking model was implemented using Pytorch Lightning[9]. We set the parameters of the FFNN

[5]Sacrebleu with max_ngram_order = 4 (https://github.com/mjpost/sacrebleu)

[6]The official implementation with roberta-large model (https://github.com/Tiiiger/bert_score)

[7]EASSE (Alva-Manchego et al., 2019)

[8]https://huggingface.co/bert-base-uncased

[9]https://www.pytorchlightning.ai

| | Simplicity-DA | | | | Human-Likert | | | | SimpDA$_{2022}$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Pearson | | Spearman | | Pearson | | Spearman | | Pearson | | Spearman | |
| | mean | std | mean | std | mean | std | mean | std | mean | std | mean | std |
| Base | .029 | .015 | .016 | .011 | .549 | .034 | .541 | .026 | .394 | .025 | .387 | .018 |
| $\overrightarrow{\text{Base}}$ (NPRM) | .012 | .030 | .000 | .013 | .434 | .057 | .451 | .042 | .308 | .047 | .336 | .021 |
| $\overleftarrow{\text{Base}}$ | .035 | .031 | .024 | .023 | .458 | .052 | .477 | .031 | .334 | .040 | .366 | .036 |
| +Silver | **.049** | .017 | **.027** | .016 | **.580** | .035 | **.547** | .033 | .366 | .026 | **.401** | .028 |
| $\overrightarrow{\text{+Silver}}$ | **.017** | .034 | **.003** | .068 | **.452** | .069 | **.483** | .051 | .255 | .049 | **.342** | .058 |
| $\overleftarrow{\text{+Silver}}$ | **.059** | .023 | **.046** | .026 | **.559*** | .060 | **.501** | .044 | .337 | .046 | **.384** | .032 |
| +Gold | .052 | .017 | .026 | .013 | .607 | .022 | .604 | .017 | .446 | .027 | .465 | .025 |
| $\overrightarrow{\text{+Gold}}$ | .025 | .020 | .019 | .020 | .535 | .038 | .561 | .027 | .393 | .039 | .421 | .026 |
| $\overleftarrow{\text{+Gold}}$ | .065 | .016 | .033 | .011 | .555 | .047 | .561 | .028 | .412 | .033 | .459 | .027 |
| SARI | .358 | - | .326 | - | .390 | - | .373 | - | - | - | - | - |
| BLEU | .507 | - | .482 | - | .349 | - | .312 | - | - | - | - | - |
| BERTScore$_p$ | .628 | - | .660 | - | .417 | - | .387 | - | - | - | - | - |
| BERTScore$_r$ | .505 | - | .502 | - | .374 | - | .401 | - | - | - | - | - |
| BERTScore$_{f1}$ | .590 | - | .579 | - | .393 | - | .393 | - | - | - | - | - |
| SAMSA | .060 | - | .068 | - | −.374 | - | −.319 | - | −.083 | - | −.122 | - |
| FKGL | .117 | - | .110 | - | −.353 | - | −.359 | - | −.387 | - | −.353 | - |

Table 1: Correlations of SIERA (top half) and other metrics (bottom half) with three evaluation datasets. The mean and standard deviation of ten runs with different seeds are shown for SIERA. The single calculation result is shown for other metrics since they have no seed. Because SimpDA$_{2022}$ has no reference, the results for the reference-based methods are not available. The bold values for the +Silver family indicate superiority over the corresponding Base value. The asterisk (*) denotes the significant difference at p < .05 of the two-sided permutation test.

| Dataset | Simplicity-DA | | Human-Likert | | SimpDA$_{2022}$ | |
|---|---|---|---|---|---|---|
| | Pearson | Spearman | Pearson | Spearman | Pearson | Spearman |
| $\Delta_{(\text{+Silver,Base})}$ | .020 | .011 | .031 | .006 | −.028 | .014 |
| $\Delta_{(\text{+Gold,+Silver})}$ | .003 | −.001 | .027 | .057 | .080 | .064 |

Table 2: Difference of the correlation coefficient (mean) between the models

and the BERT last layer learnable. AdamW[10] was chosen as the optimization algorithm, with the number of epochs set to 10 and the batch size to 16. We used cross-entropy loss[11] as the loss function and the learning rate was set to $10^{-4}$. Twenty percent of the training data was used for validation. We adopted early stopping based on the loss with the validation data and selected the checkpoint at the epoch with the lowest loss[12] We conduct the experiment with random seed values ten times, and report their average results.

**Evaluation data**

We use three data sets, Simplicity-DA, Human-Likert and SimpDA$_{2022}$, for evaluating the models. Correlations between the model prediction scores and the human ratings are calculated in two ways: Pearson's correlation coefficient and Spearman's rank correlation coefficient.

## 6.2. Results and discussion

The top half of Table 1 shows the results of the SIERA models with different settings. Comparing the two-way models (model name without an arrow) and the one-way models (those with an arrow), the two-way models are consistently superior to their one-way counterparts for Human-Likert and SimpDA$_{2022}$ but not for Simplicity-DA. Furthermore, the backwardly-ordered models are superior to the forwardly-ordered models for all datasets. We could not find an explanation for this asymmetry yet. This is an unfortunate result for NPRM, which employs the forwardly-ordered model.

+Silver outperforms Base except for the case of Pearson's coefficient with SimpDA$_{2022}$. Table 2 shows the difference between the mean values of the correlation coefficients in Table 1, where $\Delta_{(X,Y)}$

[10] https://pytorch.org/docs/stable/generated/torch.optim.AdamW.html

[11] https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html

[12] The codes are available at https://github.com/hyama1569/siera.

| | Simplicity-DA | | | | Human-Likert | | | | $\text{SimpDA}_{2022}$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Pearson | | Spearman | | Pearson | | Spearman | | Pearson | | Spearman | |
| | mean | std | mean | std | mean | std | mean | std | mean | std | mean | std |
| Base | .029 | .015 | .016 | .011 | .549 | .034 | .541 | .026 | .394 | .025 | .387 | .018 |
| +Silver(50%) | .065 | .016 | .033 | .011 | .555 | .047 | .561 | .028 | .349 | .026 | .400 | .026 |
| +Silver | .049 | .017 | .027 | .016 | .580 | .035 | .547 | .033 | .366 | .026 | .401 | .028 |
| $\Delta_{(+\text{Silver(50\%)},\text{Base})}$ | .050 | | .017 | | .006 | | .020 | | −.045 | | .013 | |
| $\Delta_{(+\text{Silver},+\text{Silver(50\%)})}$ | −.016 | | −.006 | | .025 | | −.014 | | .017 | | .001 | |
| +Gold(50%) | .039 | .018 | .015 | .017 | .601 | .032 | .593 | .032 | .439 | .028 | .459 | .019 |
| +Gold | .052 | .017 | .026 | .013 | .607 | .022 | .604 | .017 | .446 | .027 | .465 | .025 |
| $\Delta_{(+\text{Gold(50\%)},\text{Base})}$ | .010 | | −.001 | | .052 | | .052 | | .045 | | .072 | |
| $\Delta_{(+\text{Gold},+\text{Gold(50\%)})}$ | .028 | | .011 | | .006 | | .011 | | .007 | | .006 | |

Table 3: Correlations of SIERA using half of the augmented data

denotes a difference of X's value from Y's value. This result confirms the effectiveness of the proposed data augmentation method. As the augmented data used for training +Gold are made from manually written intermediate sentences by professionals, we can consider the results of +Gold as an upper bound regarding the data augmentation. Table 2 shows that the gains from Base to +Silver tend to be larger than that from +Silver to +Gold for Human-Likert and $\text{SimpDA}_{2022}$. This result suggests room for improvement in the quality of the intermediate sentences derived by applying edit operations. Although we employed the BERTScore-based filtering to exclude irrelevant intermediate sentences, this filtering is still limited. We discarded dissimilar intermediate sentences to their original and simplified sentences regarding $\text{BERTScore}_{\text{f1}}$. The similarity judgement was done against a threshold calculated by averaging the similarity of all generated intermediated sentences. The thresholds for +Silver and +Gold datasets are quite close, i.e. 0.554 and 0.547, respectively. Therefore, sentence similarity is not enough for filtering irrelevant sentences. We need to consider more effective methods for obtaining high-quality intermediate sentences.

We conducted a supplemental experiment using half of the augmented data. The result is shown in Table 3. The rows $\Delta_{(+\text{Silver},+\text{Silver(50\%)})}$ and $\Delta_{(+\text{Gold},+\text{Gold(50\%)})}$ indicate that the augmented data size reduction does not significantly impact the correlation with the human ratings. They also show that the gains from Base to +Gold(50%) are consistently larger than that from +Gold(50%) to +Gold for Human Likert and $\text{SimpDA}_{2022}$. However, this does not hold for +Silver. This difference suggests that the quality of the augmented pairs, i.e. the intermediate sentences, has more impact on the correlation than their size, supporting our claim on the importance of the intermediate sentence quality.

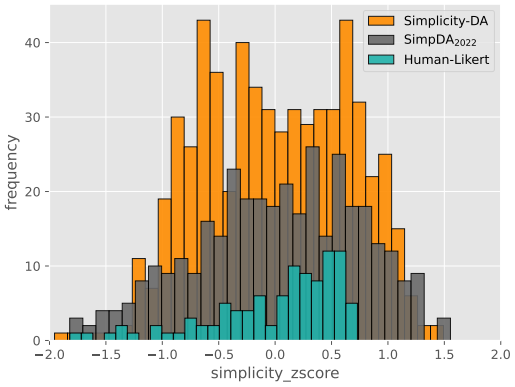


Figure 2: Distribution of simplicity scores of three datasets

The SIERA's correlation coefficients are far lower for Simplicity-DA than for the other two evaluation data sets. A possible explanation is a quality gap of simplified sentences between the training (Newsela) and evaluation data (Simplicity-DA). Simplified sentences in Newsela were written by human experts, while those in Simplicity-DA were generated by the six automatic simplification models, and four of them were rather old RNN-based models. $\text{SimpDA}_{2022}$ also includes two-thirds of its data automatically generated. However, they are all Transformer-based models, which can generate more fluent sentences than the RNN-based models used for Simplicidy-DA. We suspect that the difference between Simplicity-DA and $\text{SimpDA}_{2022}$ comes from the quality of simplified sentences to evaluate.

Figure 2 shows the distribution of human simplicity ratings of simplified sentences in the three evaluation datasets. Following Alva-Manchego et al. (2021), we normalized the ratings to z-scores ranging from −1 to 1. We notice that Simplicity-DA has a distinct lump on the left side, i.e. it has more

| | Pearson | | Spearman | |
|---|---|---|---|---|
| | mean | std | mean | std |
| Base | .236 | .023 | .264 | .027 |
| +Silver | .266 | .033 | .284 | .031 |
| +Gold | .281 | .028 | .288 | .028 |
| SARI | .121 | - | .137 | - |
| BLEU | .212 | - | .243 | - |
| $\text{BERTScore}_{\text{p}}$ | .195 | - | .203 | - |
| $\text{BERTScore}_{\text{r}}$ | .073 | - | .157 | - |
| $\text{BERTScore}_{\text{f1}}$ | .114 | - | .144 | - |
| SAMSA | .038 | - | −.005 | - |
| FKGL | −.041 | - | −.062 | - |

Table 4: Result for high-quality Simplicity-DA subset

low-rated sentences than the other two. Considering we trained the SIERA model using only human-written high-quality sentences, we suspect that the SIERA model could not learn to score low-quality simplified sentences. To confirm our hypothesis, we select high-quality simplified sentences from Simplicity-DA and calculate the correlation between their human ratings and the SIERA scores. We normalize the scores by transforming the human-rated fluency, meaning preservation and simplicity scores into z-scores, and select simplified sentences where all three scores are positive. The resultant high-quality subset contains 196 sentences, about one-third of the entire Simplicity-DA. Table 4 shows the result for the subset, which supports our hypothesis. Our above claims in Human-Likert and $\text{SimpDA}_{2022}$ also hold in the high-quality Simplicity-DA subset.

| | Pearson | | Spearman | |
|---|---|---|---|---|
| | mean | std | mean | std |
| Base | .623 | .028 | .639 | .018 |
| +Silver | .621 | .025 | .654 | .023 |
| SARI | .324 | - | .293 | - |
| BLEU | .573 | - | .511 | - |
| $\text{BERTScore}_{\text{p}}$ | .602 | - | .595 | - |
| $\text{BERTScore}_{\text{r}}$ | .507 | - | .476 | - |
| $\text{BERTScore}_{\text{f1}}$ | .578 | - | .535 | - |
| SAMSA | .078 | - | .096 | - |
| FKGL | .076 | - | .082 | - |

Table 5: Results of Simplicity-DA-trained SIERA

To further confirm our hypothesis, we trained a SIERA model using a part of the Simplicity-DA data. We randomly chose 80 simplified sentences from each simplification model of Simplicity-DA for training, resulting in 480 sentence pairs. The remaining 120 sentence pairs were held for testing. Although the model outputs could be a simplified sentence in a pair, their human rating might be very low because the models do not always work well. We define a simplified sentence in a pair based on the human simplicity rating of the model outputs. When the human rating of the model output is lower than the average rating of the entire training data, the original sentence is considered a simplified sentence. Since the number of training data was small, we increased the training data by paring system outputs from the same original sentence. The sentence with a higher human rating in each pair is considered a simplified sentence. This operation increased the training data size to 1,538 sentence pairs in total.

This training data creation refers to human ratings. This is not a normal way of training the SIERA model, which uses only pairs of original and simplified sentences without human ratings. The experiment only aims to confirm our hypothesis. Table 5 shows the result, reinforcing our hypothesis on the sentence quality gap between the training and test data.

The bottom half of Table 1, 4 and 5 shows the correlations of the other evaluation metrics. $\text{BERTScore}_{\text{p}}$ shows good performance for both Simplicity-DA and Human-Likert, which is consistent with the results of the previous study (Alva-Manchego et al., 2021). Comparing +Silver with the other evaluation metrics, we can see that +Silver consistently beats these metrics for the high-quality Simplicity-DA subset, Human-Likert and $\text{SimpDA}_{2022}$.[13] Not to mention its high correlation, SIERA has a strong point that it does not require reference simplified sentences for evaluation. As we discussed, however, we need to be careful about the training data for the SIERA model, which should be consistent with the quality of the target sentences.

## 7. Conclusion and future work

We presented SIERA, a novel reference-free metric for evaluating sentence simplicity. SIERA adopts a pair-wise ranking model to predict the order relations of simplicity in the paired sentences. The model is trained by pairs of original and simplified sentences. Evaluating simplified sentences with SIERA requires only pairs of the original and simplified sentences, i.e. reference sentences are unnecessary. We also propose a data augmentation method by applying automatically extracted edit operations to the original sentence to generate intermediate sentences. The intermediate sentences are expected to have middle simplicity between the original and corresponding simplified sentences.

We evaluated SIERA using three evaluation data sets for text simplicity. The experimental results

[13] FKGL with Pearson's coefficient is the exception.

showed that as far as the quality of target sentences is consistent with that of the training data, SIERA correlates better with human ratings than other simplicity metrics, including reference-based metrics. SIERA does not require reference sentences but needs training. We must carefully choose the training data to maximize SIERA's potential.

Also, the augmented data by the proposed method contributed to improving SIERA's correlation with human ratings. To augment the training data, we automatically extracted edit operations from a pair of the original and simplified sentences and applied a subset of the operations to the original sentence to obtain intermediate sentences. However, we did not consider dependencies among operations in the application, which may cause irrelevant sentences, as we discussed in section 5. We applied the BERTScore-based filtering to exclude irrelevant sentences, but the experimental result suggested this filtering had a limitation. Improving the quality of intermediate sentences is one of the future research directions. Considering the syntactic structure of sentences might help to generate more relevant intermediate sentences.

As we have limited parallel corpora for text simplification, we could not verify the effectiveness of metrics employing a trainable model like SIERA for different domain texts from the training data. We found that the quality gap between the training and test data impacts the performance of the SIERA model. Likewise, the domain shift would have an impact as well. Parallel corpora for simplification have been built in several domains like administrative documents (Scarton et al., 2018), general medical documents (Devaraj et al., 2021), and radiology reports (Yang et al., 2023). However, they have not necessarily been assigned human ratings. They can be used for training models but not for the evaluation of metrics. Collecting parallel corpora for simplification in various domains that can be used both for training and evaluation is indispensable.

## 8. Bibliographical References


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

Fernando Alva-Manchego, Joachim Bingel, Gustavo Paetzold, Carolina Scarton, and Lucia Specia. 2017. Learning how to simplify from explicit labeling of complex-simplified text pairs. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 295–305, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4668–4679, Online. Association for Computational Linguistics.

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.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (un)suitability of automatic evaluation metrics for text simplification. *Computational Linguistics*, 47(4):861–889.

Rémi Cardon and Adrien Bibal. 2023. On operations in automatic text simplification. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 116–130, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Rémi Cardon, Adrien Bibal, Rodrigo Wilkens, David Alfter, Magali Norré, Adeline Müller, Watrin Patrick, and Thomas François. 2022. Linguistic corpus annotation for automatic text simplification evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1842–1866, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2023. Simplicity level estimate (SLE): A learned reference-less metric for sentence simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12053–12059, Singapore. Association for Computational Linguistics.

Mohammad Dehghan, Dhruv Kumar, and Lukasz Golab. 2022. GRS: Combining generation and revision in unsupervised sentence simplification. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 949–960, Dublin, Ireland. Association for Computational Linguistics.

Ashwin Devaraj, Iain Marshall, Byron Wallace, and Junyi Jessy Li. 2021. Paragraph-level simplification of medical texts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4972–4984, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yue Dong, Zichao Li, Mehdi Rezagholizadeh, and Jackie Chi Kit Cheung. 2019. EditNTS: An neural programmer-interpreter model for sentence simplification through explicit editing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3393–3402, Florence, Italy. Association for Computational Linguistics.

Sian Gooding. 2022. On the ethical considerations of text simplification. In *Ninth Workshop on Speech and Language Processing for Assistive Technologies (SLPAT-2022)*, pages 50–57, Dublin, Ireland. Association for Computational Linguistics.

David Heineman, Yao Dou, Mounica Maddela, and Wei Xu. 2023. Dancing between success and failure: Edit-level simplification evaluation using SALSA. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3466–3495, Singapore. Association for Computational Linguistics.

J. Peter Kincaid, Robert P. Fishburne, R 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. In *Research Branch Report 8-75*.

Justin Lee and Sowmya Vajjala. 2022. A neural pairwise ranking model for readability assessment. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3802–3813, Dublin, Ireland. Association for Computational Linguistics.

Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2023. LENS: A learnable evaluation metric for text simplification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16383–16408, Toronto, Canada. Association for Computational Linguistics.

Koki Maeda, Masahiro Kaneko, and Naoaki Okazaki. 2022. IMPARA: Impact-based metric for GEC using parallel data. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3578–3588, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Louis Martin, Éric de la Clergerie, Benoît Sagot, and Antoine Bordes. 2020. Controllable sentence simplification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4689–4698, Marseille, France. European Language Resources Association.

Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2022. MUSS: Multilingual unsupervised sentence simplification by mining paraphrases. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1651–1664, Marseille, France. European Language Resources Association.

Louis Martin, Samuel Humeau, Pierre-Emmanuel Mazaré, Éric de La Clergerie, Antoine Bordes, and Benoît Sagot. 2018. Reference-less quality estimation of text simplification systems. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 29–38, Tilburg, the Netherlands. Association for Computational Linguistics.

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.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *CoRR*, abs/2203.02155.

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 on Associ-*

*ation for Computational Linguistics*, page 311–318, USA. Association for Computational Linguistics.

Colin Raffel, Noam M. 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. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2022. A survey of evaluation metrics used for NLG systems. *ACM Computing Surveys*, 55(2).

Carolina Scarton, Gustavo Paetzold, and Lucia Specia. 2018. SimPA: A sentence-level simplification corpus for the public administration domain. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Thomas Scialom, Louis Martin, Jacopo Staiano, Éric Villemonte de la Clergerie, and Benoît Sagot. 2021. Rethinking automatic evaluation in sentence simplification. *CoRR*, abs/2104.07560.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018a. BLEU is not suitable for the evaluation of text simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 738–744, Brussels, Belgium. Association for Computational Linguistics.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018b. Semantic structural evaluation for text simplification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 685–696, New Orleans, Louisiana. Association for Computational Linguistics.

Teerapaun Tanprasert and David Kauchak. 2021. Flesch-kincaid is not a text simplification evaluation metric. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 1–14, Online. Association for Computational Linguistics.

Sowmya Vajjala and Detmar Meurers. 2016. Readability-based sentence ranking for evaluating text simplification. *CoRR*, abs/1603.06009.

Hanlu Wu, Tengfei Ma, Lingfei Wu, Tariro Manyumwa, and Shouling Ji. 2020. Unsupervised reference-free summary quality evaluation via contrastive learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3612–3621, Online. Association for Computational Linguistics.

Sander Wubben, Antal van den Bosch, and Emiel Krahmer. 2012. Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1015–1024, Jeju Island, Korea. Association for Computational Linguistics.

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

Daichi Yamaguchi, Rei Miyata, Sayuka Shimada, and Satoshi Sato. 2023. Gauging the gap between human and machine text simplification through analytical evaluation of simplification strategies and errors. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 359–375, Dubrovnik, Croatia. Association for Computational Linguistics.

Ziyu Yang, Santhosh Cherian, and Slobodan Vucetic. 2023. Data augmentation for radiology report simplification. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1922–1932, Dubrovnik, Croatia. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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.

Sanqiang Zhao, Rui Meng, Daqing He, Andi Saptono, and Bambang Parmanto. 2018. Integrating transformer and paraphrase rules for sentence simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3164–3173,

Brussels, Belgium. Association for Computational Linguistics.

## 9. Language Resource References

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (un)suitability of automatic evaluation metrics for text simplification. *Computational Linguistics*, 47(4):861–889.

Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2023. LENS: A learnable evaluation metric for text simplification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16383–16408, Toronto, Canada. Association for Computational Linguistics.

Thomas Scialom, Louis Martin, Jacopo Staiano, Éric Villemonte de la Clergerie, and Benoît Sagot. 2021. Rethinking automatic evaluation in sentence simplification. *CoRR*, abs/2104.07560.

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.