Towards Reference-free Text Simplification Evaluation with a BERT Siamese Network Architecture

Xinran Zhao[1] **Esin Durmus**[1] **Dit-Yan Yeung**[2]
[1]Stanford University, [2]HKUST
{xzhaoar, esindurmus}@cs.stanford.edu, dyyeung@cse.ust.hk

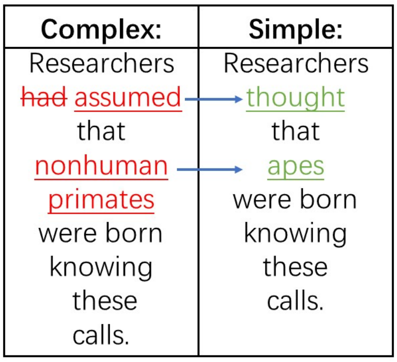
Abstract

Text simplification (TS) aims to modify sentences to make their both content and structure easier to understand. Traditional n-gram matching-based TS evaluation metrics heavily rely on the exact token match and human-annotated simplified sentences. In this paper, we present a novel neural-network-based reference-free TS metric *BETS* that leverages pre-trained contextualized language representation models and large-scale paraphrasing datasets to evaluate simplicity and meaning preservation. We show that our metric, without collecting any costly human simplification reference, correlates better than existing metrics with human judgments for the quality of both overall simplification (+7.7%) and its key aspects, i.e., comparative simplicity (+11.2%) and meaning preservation (+9.2%).

1 Introduction

Text simplification (TS) models aim at rewriting complicated input sentences into more readable variants. TS has a wide range of applications in various domains including education, language learning and journalism. It can further serve as a useful preprocessing step to simplify the downstream tasks such as parsing, machine translation, and information extraction (Chandrasekar et al., 1996). An example of TS is shown in Figure 1, where some parts of the sentence are deleted and paraphrased (e.g., *"nonhuman primates"* is paraphrased into a simpler term, *"apes"*).

The goal in TS is to simplify the sentences while retaining the original semantic meaning. To achieve this, three types of operations: *splitting, deletion, and paraphrasing* (Feng, 2008) have been widely studied in the research community. Splitting and deletion operations try to split long sentences and delete irrelevant modifiers (Angrosh et al., 2014; Siddharthan, 2006; Clarke and Lapata, 2006; Filippova et al., 2015; Rush et al., 2015). For paraphrasing, most recent work regards the TS task as a monolingual machine translation (MT) problem (Wubben et al., 2012; Narayan and Gardent, 2014; Zhang and Lapata, 2017; Nisioi et al., 2017; Zhao et al., 2020). In order evaluate text simplification, previous work uses two main kinds of metrics adapted from the metrics used in the MT literature: (1) **BLEU** (Papineni et al., 2002): the most commonly used metric for text generation which computes the exact n-gram matching between the reference and candidate; (2) **SARI** (Xu et al., 2016): an n-gram based metric which is specifically designed for TS to measure the correspondence of the preserved, deleted and added information between the system output and human-annotated simplified references.



Complex:	Simple:
Researchers ~~had~~ assumed that nonhuman primates were born knowing these calls.	Researchers thought that apes were born knowing these calls.

Figure 1: An example of TS with deletion and paraphrasing operations.

However, there are two main limitations of these metrics based on exact n-gram match between the system output and human reference (i.e., simplified sentences): (1) simple n-gram overlap count can fail to capture meaning preservation or compositional diversity (Zhang et al., 2020). Since TS systems usually rewrite word tokens to achieve simpler sentences, exact n-gram matching can not capture every variant of paraphrasing due to the diversity of human language (e.g., in Figure 1, "had assumed" can be rewritten as either "thought" or

Findings of the Association for Computational Linguistics: ACL 2023, pages 13250–13264
July 9-14, 2023 ©2023 Association for Computational Linguistics

"supposed"); (2) the heavy reliance on human reference restricts the generalization ability, where the reference can be wrong (Zhu et al., 2010) or requires experts to annotate (Xu et al., 2015). Since high-quality human annotations on sentence simplification are costly to acquire, it is hard to obtain sufficient human reference to capture the diversity in expression or to evaluate systems in a new domain.

To tackle these limitations, in this paper, we propose a reference-free evaluation metric *BETS* (**B**ERT **E**mbedding-based evaluation for **T**ext **S**implification), that leverages pre-trained contextualized language representation models (PTLMs) to improve semantic-involvement, comprehensiveness, and generalization ability. From our human evaluation on simplification quality, we observe that: (1) human perception on overall simplification quality is not identical to the simple arithmetic mean of the quality of its key factors: meaning preservation and simplicity change; (2) the quality of meaning preservation correlates poorly with the quality of simplicity change. On the other hand, the importance of these two kinds of aspectual quality can vary over different domains and purposes. Motivated by these observations, we first use contextualized language embeddings to replace the exact tokens. We then build two separate units with pre-trained neural networks to measure these aspects. Finally, we perform regression with these separate scores to acquire a flexible and optimizable metric that correlates well with human judgments on TS.

We evaluate the effectiveness of *BETS* by comparing with human annotations on both the rankings and aspectual scores (on grammar, simplicity, and meaning preservation) of TS outputs from various systems. We show that our metric correlates better with human scores than existing metrics, for each aspect. Finally, our combined metric with optimized coefficients creates a single balanced metric that can evaluate TS systems, where any of the main aspects can be further emphasized according to future applications.

2 Related Work

2.1 Text Simplification Evaluation

TS has been widely studied as a monolingual translation task solved with statistical (Zhu et al., 2010; Wubben et al., 2012; Narayan and Gardent, 2014) or neural network methods in either supervised (Nisioi et al., 2017; Zhang and Lapata, 2017; Zhao et al., 2018; Ippolito et al., 2019; Zhao et al., 2020; Kriz et al., 2019; Maddela et al., 2021) or unsupervised manner (Narayan and Gardent, 2016; Surya et al., 2019). Another line of work regards the problem as editing the input sequence with pre-defined operations and alternative tokens (Alva-Manchego et al., 2017; Dong et al., 2019; Kumar et al., 2020). Both lines of work mainly conduct evaluation with a combination of human evaluation and automatic metrics. Human annotators are generally asked to measure the fluency (or grammaticality), adequacy (or meaning preservation), and relative simplicity (lexical or structural) between the inputs and system outputs over a few test examples (around 20). In this work, besides these aspectual measurements, we also evaluate with human perception on the overall simplification quality through ranking the system outputs to avoid failure cases of score-based methods (e.g., assigning high score to copying).

The quality estimation of general sequence-to-sequence problems (Papineni et al., 2002; Martins et al., 2017; Specia et al., 2018; Fonseca et al., 2019; Xenouleas et al., 2019), as well as for TS (Sulem et al., 2018a), is another long-lasting research direction in the community. To design metrics specialized for TS, Xu et al. (2016) proposed two n-gram matching-based light-weight metrics. The first one, FKBLEU, is a combination of the Flesh-Kincaid index (FK) (Kincaid et al., 1975) (as readability measurement) and iBLEU (Sun and Zhou, 2012).The other one, SARI, measures if the output deletes, keeps, or adds the same n-grams as the operations of human reference sentences on the original inputs. With a different focus, SAMSA (Sulem et al., 2018b) evaluates sentence-level simplification such as splitting complex sentences. None-neural-based reference-free TS metrics have also been explored in Martin et al. (2018); Kriz et al. (2020). In our work, we propose to use a neural-network-based method to improve the measurement of simplicity change and relieve the reliance of reference (at least one reference is needed for each test sentence for n-gram matching-based methods) at the same time.

More recently, Alva-Manchego et al. (2021) calls for TS metrics that emphasize the relation between input/output and allow personalizing. *BETS* leverages contextualized embeddings to measure directly with input/output pair and computes aspectual scores to improve personalizing weights in the final score on these aspects.

2.2 Automated Evaluation with Embeddings

A large body of literature has explored the possibility of using learned dense token representations (i.e., embeddings) to capture the semantics on word level or sentence level. (e.g., word2vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), and BERT (Devlin et al., 2019)). Various approaches have been developed to apply embeddings in the evaluation of sequence-to-sequence problems to capture the lexical and structural similarity between sentences (Kusner et al., 2015; Servan et al., 2016; Lo, 2017; Tättar and Fishel, 2017; Lo et al., 2018; Chow et al., 2019; Clark et al., 2019). Recently, pre-trained contextualized language representation models (Devlin et al., 2019) have demonstrated significant improvement in capturing the contextualized word semantics. To leverage this property, Zhang et al. (2020) proposed replacing n-gram matching with greedy embedding similarity matching. Maynez et al. (2020) and Durmus et al. (2020) proposed to probe knowledge from the contextualized embedding based question answering models to evaluate the faithfulness of neural generation models. Following this line of work, we use pre-trained embeddings to capture relative readability and semantic similarity.

2.3 Automated Evaluation with Optimization

Researchers also explore ways to use learned metrics to approach human judgments by performing regression over n-grams (Stanojević and Sima'an, 2014), embeddings (Ma et al., 2017), or different embedding models (Shimanaka et al., 2018). Unlike these models which optimize the correlation with human annotations on the targeted datasets, our regression model is solely trained to capture the relation between the overall simplification quality and the performance on its key aspects (i.e., meaning preservation and comparative simplicity).

3 Methodology

3.1 Task Description

The task of TS requires a model to generate a more readable sentence that preserves the meaning of the original input. We aim to evaluate the overall simplification quality, as well as two key problems of this task: comparative simplicity and meaning preservation. Intuitively, too much simplification may hurt meaning preservation, vice versa. Accurate measurement on one aspect may not transfer to the other. As a result, we propose to evaluate text simplification with a parametric combined metric on separate components measuring these two key aspects.

3.2 Parametric combination

Previous metrics usually use a single metric to approximate the human perception on both comparative simplicity and meaning preservation. They then use the average of these aspectual human annotations to represent the overall simplification performance. However, there are two problems with this setting:

1. Intuitively, comparative simplicity and meaning preservation are not highly correlated. For example, keeping all the original tokens achieves perfect meaning preservation but no comparative simplicity. On the other hand, keeping only the main narrative simplifies the sentence on the cost of hurting meaning preservation.

2. Human perception on the TS quality may not be the simple arithmetic mean over the aspectual scores. For example, humans can be more sensitive to comparative simplicity when the meaning preservation quality surpasses a certain cognitive threshold but less sensitive when too little meaning is preserved. Moreover, different purposes and domains may require different combination of these aspects to be taken into account.

Above intuition motivates us to use a parametric metric to capture the overall performance of TS. We denote the output scores for the comparative simplicity component and the meaning preservation component as P_{simp} and $R_{meaning}$, respectively. The overall score S is a parametric combination that can be written as:

$$S = \alpha P_{simp} + \beta R_{meaning}$$

,where α and β are hyperparameters to be tuned.

3.3 Comparative Simplicity Measurement

The Flesch-Kincaid index (FK) (Kincaid et al., 1975) has long been accepted as a measure of sentence readability but suffers from indirect measurement [1] and limited training data size and domain(i.e., 531 Navy personnel manuals only).

[1] It defines simplicity as the weighted sum of word count per sentence and syllables count per word.

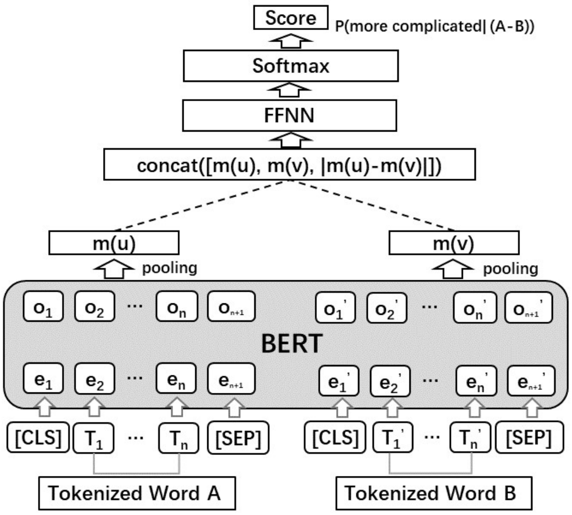


Figure 2: The siamese network structure of the comparative simplicity measurement model. The input can be either words or phrases. The output denotes if Word A is more complicated than Word B.

Name	Example
Simple PPDB	destabilise → destabilize: 0.505 resolve → solve: 0.997 phones → telephones: 0.345
Simple PPDB++	destabilise → destabilize: 0.481299 (no-diff) resolve → solve: 0.909 (simplifying) phones → telephones: -0.720 (complicating)
SemEval 2012	When you think about it, that's pretty terrible. **Alternatives** (easy→hard): 1.bad 2.awful 3.deplorable

Table 1: Examples from the involved datasets. *SemEval 2012* denotes its English Lexical Simplification shared-task. *Alternatives* denotes the candidates that can replace the underlined word, and can be in tied rank in terms of hardness.

To achieve a direct and general approximation over word readability, we leverage a large-scale dataset of simplification rules. SimplePPDB++ (Maddela and Xu, 2018) is a dataset with over 10 million rules on simplifying paraphrase rules built upon SimplePPDB (Pavlick and Callison-Burch, 2016) with human annotations on readability. Motivated by Maddela and Xu (2018); Reimers and Gurevych (2019), we propose to create a phrase-level simplicity comparison model with a BERT siamese network structure optimized by these rules in English. The generation of rules only requires workers on Amazon Mechanical Turk to do classification on automatically mined rules or common words (for SimplePPDB++), which is simpler and clearer compared to asking them to write simplified sentences (i.e., reference).

For a new domain, our trained model that captures domain-agnostic comparative similarity can be applied directly without collecting references. For further specialization, take the medical domain as an example, only the new terms need to be annotated to prepare the data for pre-training. For a new language, the formulation of PPDB is language-agnostic, which relieves the cost of finding high-quality annotators to simplify the sentences in the language (e.g., annotators are experienced teachers for Newsela (Xu et al., 2015)).

We train the model with 11,996 rules from SimplePPDB++. We pre-process the scores from -1 to 1 with thresholds ±0.4 as in original work. We get 3989, 4029, 3978 pairs labeled simplifying, complicating, or no-difference, respectively. Figure 2 shows the structure of the complexity comparison model. We perform mean pooling to get unified-dimension vectors to represent the words/phrases.

We intermediately evaluate the effectiveness of the simplicity comparison model with the English Lexical Simplification shared-task at SemEval 2012 (Specia et al., 2012). It contains 201 target words and 10 context sentences for each. To test the generalizability of the model, we conduct an unsupervised setting that applies our model directly to the original test set of the dataset with 1,710 sentences, where each sentence has 6 alternatives for the target words on average. As a result, the fine-tuning process applied to our model presents significant improvement (from 18.2% to 38.44%) over the original BERT weights (BERT-base) on predicting the simplest candidate (P@1), which demonstrates that the training process largely improves the model's ability to measure the comparative simplicity. Examples from these involved datasets are presented in Table 1.

With the simplicity comparison model (model output is denoted as $f(a, b)$) in hand, we define the sentence-level comparative simplicity as the averaged simplicity change for the added/paraphrased tokens in the output sentence. For each word in the input u and output v, with $m(x)$ as the embedding function (BERT output in Figure 2), we acquire the pooled embeddings $m(u_i)$ and $m(v_j)$. For the unique words $v_j \in v \setminus u$ (potentially simplified expressions), we find the most similar $u_i^{(j)}$ (by cosine similarity $cos(\cdot)$) and compute their related complexity. The comparative simplicity score P_{simp} is defined as the average of the pair-wise scores:

$$u_i^{(j)} = \arg\max_{u_i \in u} cos(\boldsymbol{m}(u_i), \boldsymbol{m}(v_j)) \quad (1)$$

$$P_{simp} = \frac{1}{|v \setminus u|} \sum_{v_j \in v \setminus u} f\left(u_i^{(j)}, v_j\right) \quad (2)$$

3.4 Meaning Preservation Measurement

The measurement of meaning preservation focuses on how much of the original content is preserved in the simplified variant, where the simplicity change is already covered in Section 3.3. Since the deletion operation can largely change grammar correctness, completeness, and semantics, we evaluate the semantic similarity between the input and output for all the words in the input. For each word other than the stop words (e.g., the, a, an) in the input u and output v, we acquire the pooled embeddings $m(u_i)$ and $m(v_j)$. We greedily match the words in the two sentences with a word-level maximum cosine similarity ($cos(\cdot)$) score. We find the most similar v_j and compute their similarity. The overall meaning preservation score $R_{meaning}$ is computed as a pair-wise average:

$$R_{meaning} = \frac{1}{|u|} \sum_{u_i \in u} \max_{v_j \in v} cos(\boldsymbol{m}(u_i), \boldsymbol{m}(v_j)) \quad (3)$$

3.5 Finding Coefficients

To acquire a generalizable and high-quality set of coefficients, we propose to leverage the high-quality human-written simplified sentences (i.e., gold reference) and commonly used adversarial noise in neural machine translation systems to form minimum supervision. With the original sentences and the human simplification reference as the initial step, we can contrastively form two labels: original → reference: 1 (good simplification); reference → original : -1 (bad simplification). Following the noise addition steps for neural machine translation (Lample et al., 2017, 2018; Févry and Phang, 2018), we get four kinds of adversarial examples: dropping, adding, shuffling, and substituting tokens, with details introduced in Appendix. With 289 sentences and 8 gold references[2] for each, we generate 2,601 positive, 289 neutral, and 1,156 negative examples. We then use a feature-based logistic regression model to optimize α and β. As a result, one possible set of α, β is 0.508, 2.944 respectively. We also define a vanilla version with α, β set as 0.5, 0.5 (after re-scaling P_{simp} and $R_{meaning}$ to the same range). There is potential to further improve the metric with fine-grained human annotations (e.g., (Sellam et al., 2020)). However, we only use automatic adversarial signals as minimum self- supervision to keep the metric lightweight.

4 Experiments

4.1 Experiment Setup

To compare the automatic metrics, we compute the correlation between the metric scores and human perception. Following previous work, we evaluate correlation with two kinds of evaluation protocols:

1. **System Quality**: We create our survey by adapting the standard evaluation survey proposed in Sulem et al. (2018c). We first select 70 sentences from the dataset as previous work (Xu et al., 2016) and collect the corresponding outputs from various systems (described in Section 4.2). We then ask three fluent English speakers to rate the system outputs on four parameters: Grammatical Correctness (G), Meaning Preservation (M), Simplicity (S), Structural Simplicity (StS), with a 5-point Likert Scale (Strongly agree, Agree, No opinion, Disagree, and Strongly disagree) [3]. Grammatical Correctness (G) and Meaning Preservation (M) are also referred to as Fluency and Adequacy in some previous work. The overall sentence simplicity (S+) can be represented by the average of S and StS. The average human score (denoted as *AvgHuman*) is computed as the average of G, M, and S+. In addition to our collection, we check the transferability of our metric with Simplicity-DA (Alva-Manchego et al., 2021) dataset, which collects human annotated overall simplification scores for 100 sentence-simplification pairs.

2. **System Ranking**: Most previous work represents the overall quality of each candidate sentence with an average of the annotated scores from different perspectives (i.e., G, M, S+). However, the arithmetic mean can fail to capture some unfaithful cases (e.g., copying the input) and it may not represent human perception of the overall quality. Table 2 shows an example of good simplification ties with a failed copy

[2] We used the high-quality annotations shared in github.com/cocoxu/simplification.

[3] The details and examples of the aspects are mentioned in the survey guidelines.

Input	Today NRC is organised as an independent, private foundation .
System 1 Output	Today NRC is organised as an independent, private foundation. (G,M,S+ = 5, 5, 0)
System 2 Output	Today NRC is organize as an open and private trust. (G,M,S+ = 3, 4, 3)

Table 2: An example of the tied system outputs rated by the average of human annotations. While the first output makes no change, the second output simplifies both the structure and lexicons.

due to some minor grammatical errors and information loss. This indicates the importance of introducing a ranking test that directly evaluates the human perception on the simplification quality. The annotators are provided with outputs from all the systems to compare them directly. In this setting, similar to the first protocol, for each of the 70 selected input sentences, we ask the human annotators to rank the quality of their simplified versions from the systems.

4.2 Metric and System Selection

We compare our metric with other most commonly used metrics for TS, including: (1) Flesh-Kincaid Grade Level (FK) (Kincaid et al., 1975), which computes the text readability with the number of cognitive steps needed. Since higher scores indicate lower readability, we report its opposite number -FK. (2) BLEU (Papineni et al., 2002), one of the most widely used n-gram based metrics measuring the quality of text generation. (3) iBLEU (Sun and Zhou, 2012) and FKBLEU (Xu et al., 2016), two variants of TS-specified BLEU through involving human references. We set all hyperparameters as the original work. (4) SARI (Xu et al., 2016), an n-gram matching-based metric that evaluates the operation correspondence with human reference in adding, deleting, and keeping tokens. Three component scores, F_{add}, F_{keep}, and P_{del}, are proposed to evaluate these operations respectively. The overall score is given by the arithmetic mean of the above n-gram matching scores. (5) BERTScore (Zhang et al., 2020), a powerful automatic metric for general text generation, which computes the token similarity using contextual embedding. Similarly as BLEU, we compute the similarity between system outputs and human references with BERTScore to use it as a reference-based TS metric.

For each of the metrics, we follow the original work to decide the inclusion of the single and multiple reference settings. In the multiple reference setting, we use 8 Amazon Turker annotations collected by Xu et al. (2016). Since our metric mainly focuses on sentence-level rewriting instead of inter-sentence splitting, we do not compare with SAMSA (Sulem et al., 2018b). However, for the evaluation on Simplicity-DA, we additionally include SAMSA as the original work.

To ensure a broad coverage on TS system selection, we collect the text simplification outputs from both the widely compared classic systems and the state-of-the-art (SOTA) systems: (1) PBMT (Wubben et al., 2012), which treats TS as a monolingual machine translation problem; (2) Hybrid (Narayan and Gardent, 2014), which leverages a sentence semantic tree and a machine translation system together for TS; (3) Dress-LS (Zhang and Lapata, 2017), a reinforcement learning based model for TS, where LS indicates the involvement of lexicon simplification processing (denoted as DRESS); (4) UNTS (Surya et al., 2019), an unsupervised model which does not require aligned data; (5) Edit-Unsup-TS (Kumar et al., 2020), a phrase-level editing system which improves the controllability and interpretability (denoted as EditUTS); (6) BTRLTS and BTTS (Zhao et al., 2020), SOTA unsupervised and semi-supervised systems using back-translation and denoising autoencoders. For systems with multiple variants, we select the best-performing variant reported. We collect the system outputs from Alva-Manchego et al. (2019).

4.3 Annotation Results

In total, we have collected 560 system outputs from 8 systems and 70 sentences in the PWKP test dataset. Then, we collect the evaluation of the output quality and overall ranking of the systems.

To measure the quality of the collected annotation, for system quality annotation, we estimate the inter-annotator agreement (IAA) with both the average pairwise absolute agreement (i.e., # of matched annotations / # of annotations) and Cohen's weighted Kappa (Cohen, 1968). The overall agreement with all systems and metrics is 0.69 (for pairwise absolute agreement) and 0.94 (for quadratic weighted kappa). The agreement for each system and metric pair is shown in Appendix. Such good agreement for a question with 5 options shows that the annotators can understand and solve the tasks well. The comparatively lower agreement on evaluating simplicity (i.e., S and StS) also match

Metric	ref.	G	M	S+
-FK	none	0.155	0.162	0.002*
BLEU	single	0.375	0.475	0.068*
BLEU	multiple	0.605	0.666	0.067*
iBLEU	single	0.321	0.406	0.070*
iBLEU	multiple	0.600	0.647	0.071*
FKBLEU	multiple	0.459	0.684	-0.091
SARI	single	0.063 *	0.002*	0.224
SARI	multiple	0.277	0.205	0.316
SARI – F$_{add}$	multiple	0.156	0.150	0.304
SARI – F$_{keep}$	multiple	0.642	0.752	0.051*
SARI – P$_{del}$	multiple	-0.319	-0.497	0.228
BERTScore	multiple	0.543	0.539	0.093
P_{simp}	none	0.061*	0.012*	**0.351**
$R_{meaning}$	none	**0.714**	**0.831**	0.097

Table 3: The metric correlation with human annotations with Pearson correlation. We use * to denote entries with p-value >0.05. The ref. column denotes the number of references used (8 vs 1). Avg. represents the average human scores calculated in Section 4.1. P_{simp}, $R_{meaning}$ denote the component scores in our metric that measure the simplicity (i.e., $\alpha = 1$, $\beta = 0$) and meaning preservation (i.e., $\alpha = 0$, $\beta = 1$), respectively.

the findings from the annotation process in previous work (Sulem et al., 2018b).

To test our assumption in Section 3.5, we compare the meaning preservation (M) annotations and simplicity (S) or structural simplicity (StS) annotations. The Pearson correlation scores are 0.22 and 0.05, which indicates a poor correlation between these aspects of TS and supports our intuition to assess these aspects separately.

For system ranking annotations, we calculate the average pairwise agreement and Cohen's weighted Kappa among humans by splitting the rank into comparison over each pair of entries. We conduct ordered comparisons for each annotator pair and each input sentence. The average absolute agreement between humans is 0.70. In contrast, if we use the average score of all aspects (G, M, S+) to extract the ranks, the absolute agreement between the calculated ranks and human-annotated ranks is 0.42 (-40%), which suggests the necessity of using both aspectual scores and ranking to evaluate.

5 Results

5.1 Component Metric Performance

In this section, we first evaluate how our proposed component metrics correlate with human perception for different aspects (i.e., S for comparative simplicity measurement P_{simp} and G, M for meaning preservation measurement component $R_{meaning}$). Table 3 presents the correlation of human perceptions with our component metric and baselines. From the table we can observe that, without the need for human references, the component scores P_{simp} and $R_{meaning}$ present the best correlation on these aspects aspects (+11.2% in S and +9.2% in M, respectively, comparing to the previous best). Our results also show similar findings as in the previous work (Xu et al., 2016; Sulem et al., 2018a):

1. Although BLEU and its variants significantly correlate with grammar correctness and meaning preservation annotations, they have low or even negative correlation with comparative simplicity, which limits their use in TS evaluation.

2. Although the SARI score and its components achieve a much better correlation with comparative simplicity than BLEU, its lower correlation with grammar and meaning preservation leads to a low correlation on the average .

3. The metric performance on all these n-gram based scores largely depends on the number of references. The correlation drops drastically with only one reference, which signifies of the cost issue to generalize these metrics to new datasets and domains since multiple human written references are required for each new example in the test set. In contrast, our metric achieves better correlation without relying on any reference.

5.2 Overall Correlation

Table 4 shows the correlation between automatic metrics and human judgment for system quality score (i.e., average of G, M and S+) and system rankings. Similarly, our results are consistent with findings in Section 5.1: single reference hurts the overall performance of n-gram matching-based metric. These observations match our assumptions on the challenge of using a single reference in TS evaluation.

For quality score correlation, *BETS* (regression) achieves the best performance with 7.7% and 7.4% improvement on quality correlation and ranking correlation than the second best, respectively. Self-supervised regression also helps improve *BETS* (regression), comparing to *BETS* (vanilla)

For the correlation with rankings, with a similar setting to Section 4.3, we compute the Pearson cor-

Metric	ref.	Quality Corr.	Rank Corr.
Avg. Human	none	0.941	0.664
BLEU	single	0.419	0.325
BLEU	multiple	0.662	0.514
iBLEU	single	0.361	0.238
iBLEU	multiple	0.655	0.432
FKBLEU	multiple	0.535	0.544
SARI	single	0.066*	0.076*
SARI	multiple	0.297	0.343
SARI – F_{add}	multiple	0.219	0.230
SARI – F_{keep}	multiple	0.684	0.548
SARI – P_{del}	multiple	-0.350	-0.041*
BERTScore	multiple	0.562	0.590
BETS (vanilla)	none	0.466	0.451
BETS (regression)	none	**0.737**	**0.634**

Table 4: The metric correlation with human quality score and rankings with Pearson correlation. We use * to denote entries with p-value >0.05. *BETS* (vanilla) denotes using $\alpha = 0.5$ and $\beta = 0.5$. *BETS* (regression) denotes using coefficients optimized from minimum-supervised regression. The ref. and Corr. denote the number of human references used and correlation, respectively. Best performed entries is marked in bold.

relation on all the pair-wise ranks (i.e., "equal or better" and "worse") between the metric scores and human judgments on the ranking. In general, our metrics show a good correlation with human rankings, which suggests that these metrics show robustness over possibly unfaithful errors (e.g., copying).

In practice, with our parametric metric, the *optimal* point can be adjusted through adjusting component weights for different applications.

5.3 External evaluation: Simplicity-DA

Similar findings can be observed from from Table 5 on Simplicity-DA (Alva-Manchego et al., 2021), which contains human judgments of simplification quality on TS outputs from six systems elicited via direct assessment (Graham et al., 2015) from Amazon Mechanical Turkers. *BETS* (regression) outperforms other none-reference-based metrics by a large margin and achieves the second best correlation (only lower than our $R_{meaning}$ component metric) among all the metrics without any reference or further optimization on Simplicity-DA, which suggests this metric is generalizable to different datasets.

An interesting finding is that BLEU performs well and outperforms SARI on this dataset, which differs from earlier findings in Sulem et al. (2018a). One possible reason is that copying original sentences may not get penalized during the annotation process of Simplicity-DA, as stated in the original work. As a result, meaning preservation can dominate the score in these cases. Therefore, text generation evaluation metrics (BLEU, BERTScore, and *BETS*-$R_{meaning}$) may get higher correlations.

Metric	ref.	Corr.
BLEU	multiple	0.405
iBLEU	multiple	0.398
FKBLEU	multiple	0.131
SARI	multiple	0.336
BERTScore	multiple	0.518
-FK	none	0.272
SAMSA	none	0.103
BETS - P_{simp}	none	0.079
BETS - $R_{meaning}$	none	**0.755**
BETS (vanilla)	none	0.254
BETS (regression)	none	0.618

Table 5: The metric Pearson correlation with Simplicity-DA scores (p-value is not included in the original work). The notations for the variants of *BETS* are the same as Table 4. Best performed entry is marked in bold.

6 Conclusion

In this paper, we investigate the problem of TS evaluation. We find that: (1) different key aspects of TS, comparative simplicity and meaning preservation, correlate poorly and should be measured separately; (2) the overall simplification quality should not be solely evaluated with arithmetic mean of the scores for these aspects. Other human annotation formats, such as ranking, should also be considered to capture diverse human perception.

Upon such findings, we propose to leverage large-scale simplification rule bases and PTLMs to evaluate TS. We build two component metrics focusing on the aforementioned key aspects. These reference-free metrics correlate better with human judgment on their specialized aspects than existing metrics that require multiple human reference sentences. This allows our metrics to be able to transfer to new tasks and domains. We further combine these metrics with optimizable coefficients to create a balanced general metric on all the aspects.

Experiments show that our metric, *BETS*, correlates well with human judgments in both the scoring and ranking settings. Also, the focus on preservation or simplifying text can be manually adjusted by changing the weights on two almost orthogonal components to fit personalized and specific weightings of these aspect in different domains.

7 Acknowledgement

This research has been made possible by funding support from the Research Grants Council of Hong Kong under the General Research Fund project 16204720. We are grateful to Hongming Zhang, Shikhar Murty, Tianqing Fang, and the anonymous reviewers for their helpful comments.

8 Reproduciblity

Our code is available at: `https://github.com/colinzhaoust/reference-free_TS_evaluation`.

9 Limitations

Under-explored multilingual generalizability. We evaluate the efficiency of our metric mainly for English. We aim to extend this work to other languages using language-agnostic PPDB as described in Section 3.2.

Restricted to lexical simplicity. We mainly examine the lexicon-level effect of the quality of the text simplification (i.e., the relative simplicity and meaning preservation of the output tokens). Other sentence-level factors that could have an effect in simplicity is not explored in this paper, such as the compositional difficulty (e.g., whether the sentence uses a inverted order) and comprehension difficulty (e.g., a sentence written with simpler words may still be hard to understand). Empirically, we observe that all the involved metrics correlate poorly with StS. We aim to conduct further research exploring the automatic metrics on sentence-level simplicity without external knowledge bases or human references.

References

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, 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*, pages 1–29.

Mandya Angrosh, Tadashi Nomoto, and Advaith Siddharthan. 2014. Lexico-syntactic text simplification and compression with typed dependencies. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1996–2006, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

R. Chandrasekar, Christine Doran, and B. Srinivas. 1996. Motivations and methods for text simplification. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*.

Julian Chow, Lucia Specia, and Pranava Madhyastha. 2019. WMDO: Fluency-based word mover's distance for machine translation evaluation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 494–500, Florence, Italy. Association for Computational Linguistics.

Elizabeth Clark, Asli Celikyilmaz, and Noah A. Smith. 2019. Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2748–2760, Florence, Italy. Association for Computational Linguistics.

James Clarke and Mirella Lapata. 2006. Models for sentence compression: A comparison across domains, training requirements and evaluation measures. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 377–384, Sydney, Australia. Association for Computational Linguistics.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186.

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.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Lijun Feng. 2008. Text simplification: A survey. *The City University of New York, Technical Report*.

Thibault Févry and Jason Phang. 2018. Unsupervised sentence compression using denoising autoencoders. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 413–422, Brussels, Belgium. Association for Computational Linguistics.

Katja Filippova, Enrique Alfonseca, Carlos A. Colmenares, Lukasz Kaiser, and Oriol Vinyals. 2015. Sentence compression by deletion with LSTMs. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 360–368, Lisbon, Portugal. Association for Computational Linguistics.

Erick Fonseca, Lisa Yankovskaya, André F. T. Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the WMT 2019 shared tasks on quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 1–10, Florence, Italy. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2015. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23:3 – 30.

Daphne Ippolito, Reno Kriz, Joao Sedoc, Maria Kustikova, and Chris Callison-Burch. 2019. Comparison of diverse decoding methods from conditional language models. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics.

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.

Reno Kriz, Marianna Apidianaki, and Chris Callison-Burch. 2020. Simple-qe: Better automatic quality estimation for text simplification. *CoRR*, abs/2012.12382.

Reno Kriz, João Sedoc, Marianna Apidianaki, Carolina Zheng, Gaurav Kumar, Eleni Miltsakaki, and Chris Callison-Burch. 2019. Complexity-weighted loss and diverse reranking for sentence simplification. *CoRR*, abs/1904.02767.

Dhruv Kumar, Lili Mou, Lukasz Golab, and Olga Vechtomova. 2020. Iterative edit-based unsupervised sentence simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7918–7928, Online. Association for Computational Linguistics.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 957–966, Lille, France. PMLR.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. In *Proceedings of EMNLP 2018*, pages 5039–5049, Brussels, Belgium. Association for Computational Linguistics.

Chi-kiu Lo. 2017. MEANT 2.0: Accurate semantic MT evaluation for any output language. In *Proceedings of the Second Conference on Machine Translation*, pages 589–597, Copenhagen, Denmark. Association for Computational Linguistics.

Chi-kiu Lo, Michel Simard, Darlene Stewart, Samuel Larkin, Cyril Goutte, and Patrick Littell. 2018. Accurate semantic textual similarity for cleaning noisy parallel corpora using semantic machine translation evaluation metric: The NRC supervised submissions to the parallel corpus filtering task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 908–916, Belgium, Brussels. Association for Computational Linguistics.

Qingsong Ma, Yvette Graham, Shugen Wang, and Qun Liu. 2017. Blend: a novel combined MT metric based on direct assessment — CASICT-DCU submission to WMT17 metrics task. In *Proceedings of the Second Conference on Machine Translation*, pages 598–603, Copenhagen, Denmark. Association for Computational Linguistics.

Mounica Maddela, Fernando Alva-Manchego, and Wei Xu. 2021. Controllable text simplification with explicit paraphrasing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3536–3553, Online. Association for Computational Linguistics.

Mounica Maddela and Wei Xu. 2018. A word-complexity lexicon and a neural readability ranking model for lexical simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3749–3760, Brussels, Belgium. Association for Computational Linguistics.

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.

André F. T. Martins, Marcin Junczys-Dowmunt, Fabio N. Kepler, Ramón Astudillo, Chris Hokamp, and Roman Grundkiewicz. 2017. Pushing the limits of translation quality estimation. *Transactions of the Association for Computational Linguistics*, 5:205–218.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26, pages 3111–3119. Curran Associates, Inc.

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.

Shashi Narayan and Claire Gardent. 2016. Unsupervised sentence simplification using deep semantics. In *Proceedings of the 9th International Natural Language Generation conference*, pages 111–120, Edinburgh, UK. Association for Computational Linguistics.

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 85–91, Vancouver, Canada. 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 on association for computational linguistics*, pages 311–318.

Ellie Pavlick and Chris Callison-Burch. 2016. Simple PPDB: A paraphrase database for simplification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 143–148, Berlin, Germany. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Christophe Servan, Alexandre Bérard, Zied Elloumi, Hervé Blanchon, and Laurent Besacier. 2016. Word2Vec vs DBnary: Augmenting METEOR using vector representations or lexical resources? In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1159–1168, Osaka, Japan. The COLING 2016 Organizing Committee.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2018. RUSE: Regressor using sentence embeddings for automatic machine translation evaluation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 751–758, Belgium, Brussels. Association for Computational Linguistics.

Advaith Siddharthan. 2006. Syntactic simplification and text cohesion. *Research on Language and Computation*, 4(1):77–109.

Lucia Specia, Frédéric Blain, Varvara Logacheva, Ramón F. Astudillo, and André F. T. Martins. 2018. Findings of the WMT 2018 shared task on quality estimation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 689–709, Belgium, Brussels. Association for Computational Linguistics.

Lucia Specia, Sujay Kumar Jauhar, and Rada Mihalcea. 2012. SemEval-2012 task 1: English lexical simplification. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of SemEval 2012*, pages 347–355, Montréal, Canada. Association for Computational Linguistics.

Miloš Stanojević and Khalil Sima'an. 2014. BEER: BEtter evaluation as ranking. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 414–419, Baltimore, Maryland, USA. Association for Computational Linguistics.

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.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018c. Simple and effective text simplification using semantic and neural methods. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 162–173, Melbourne, Australia. Association for Computational Linguistics.

Hong Sun and Ming Zhou. 2012. Joint learning of a dual SMT system for paraphrase generation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 38–42, Jeju Island, Korea. Association for Computational Linguistics.

Sai Surya, Abhijit Mishra, Anirban Laha, Parag Jain, and Karthik Sankaranarayanan. 2019. Unsupervised neural text simplification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2058–2068, Florence, Italy. Association for Computational Linguistics.

Andre Tättar and Mark Fishel. 2017. bleu2vec: the painfully familiar metric on continuous vector space steroids. In *Proceedings of the Second Conference on Machine Translation*, pages 619–622, Copenhagen, Denmark. 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.

Stratos Xenouleas, Prodromos Malakasiotis, Marianna Apidianaki, and Ion Androutsopoulos. 2019. SUM-QE: a BERT-based summary quality estimation model. 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)*, pages 6005–6011, Hong Kong, China. Association for Computational Linguistics.

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. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

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.

Yanbin Zhao, Lu Chen, Zhi Chen, and Kai Yu. 2020. Semi-supervised text simplification with back-translation and asymmetric denoising autoencoders. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:9668–9675.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1353–1361, Beijing, China. Coling 2010 Organizing Committee.

A Appendix

A.1 Implementation Details

We conduct our experiments on 4 GTX 1080 Ti graphics cards with CUDA 11 installed.

We initialize the comparative simplicity measurement model using the BERT-base model with 110M parameters. We pad the input tokens/sentences with BERT start and end symbols (i.e., [CLS] and [SEP]). The size of the hidden states for *FFNN* blocks is 256. During training, we use Mean Squared Error (MSE) loss as the loss function and stochastic gradient descent (SGD) as the optimizer. We initialize all parameters randomly and train all the models with 3 epochs. On average, each epoch takes 20 hours. We use the SemEval 2012 English Lexical Simplification shared task as the intermediate evaluation. We use precision@1 and Pearson correlation to evaluate the models. We tune the hyperparameters with uniform sampling for 10 times and set the learning rate as 5e-5 and the maximum phrase length as 15. The results are presented in the main paper. Other component metrics (P_{simp} and $R_{meaning}$) use pretrained weights from the comparative simplicity measurement model and BERT-base model with 110M parameters. We use the Scikit-learn package [4] to compute logistic regression. All parameters are set as default.

A.2 Details of finding coefficients

The details of each way to find adversarial examples are introduced in Table 6. Besides substitution, all methods lead to examples with negative examples. A figurative illustration on the effect of involving these noise types can be found in Figure 3, where adequacy and simplicity denote the quality of meaning preservation and comparative simplicity change, respectively.

A.3 Annotation Quality

Table 7 presents the scores for system quality annotation. The overall agreement with all systems and metrics is 0.69 (for pairwise absolute agreement) and 0.94 (for quadratic weighted kappa). Such good agreement for a question with 5 options shows that the annotators can understand and solve the tasks well. The quality of the agreement and comparatively lower agreement on evaluating simplicity (i.e., S and StS) also match the annotation results from previous work (Sulem et al., 2018b).

Noise	Description (label)
Substitution	We use the substitution rules from SimplePPDB or SimplePPDB++ to replace the tokens in the input. (simplifying rule: 1; complicating rule: -1; no-difference rule: 0)
Drop	We drop the tokens from the gold reference at a certain probability. (-1)
Additive	We sample a subsequence from another sentence in the corpus and append it at the end of the original sentence. (-1)
Shuffling	We shuffle the original sentence to break the original semantics. (-1)

Table 6: Involved noise types and their descriptions for adversarial example generation. $1, -1, 0$ denote the process to create good/bad/no-difference simplification, respectively.

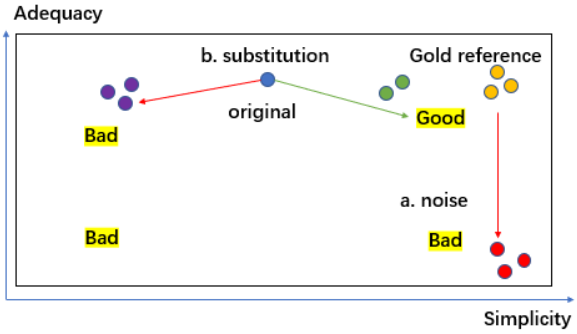


Figure 3: By adding noise (e.g., shuffling, deletion) to the human written reference (gold reference) and substitution on original sentences, we can acquire adversarial examples with different characteristics: high adequacy, low simplicity, or vice versa.

Model	G	M	S	StS
Overall	0.70(0.63)	0.63(0.75)	0.77(0.41)	0.65(0.27)
Reference	0.81(0.24)	0.73(0.59)	0.61(0.49)	0.57(0.48)
Dress	0.91(0.58)	0.74(0.86)	0.86(0.28)	0.54(0.26)
Hybrid	0.46(0.58)	0.53(0.68)	0.93(0.27)	0.29(-0.05)
PBMT	0.77(0.53)	0.61(0.55)	0.66(0.44)	0.74(0.30)
UNTS	0.63(0.69)	0.61(0.79)	0.79(0.22)	0.74(0.12)
EditUTS	0.60(0.47)	0.50(0.42)	0.71(0.18)	0.71(0.33)
BRTLTS	0.69(0.34)	0.63(0.70)	0.77(0.32)	0.84(0.42)
BTTS10	0.76(0.74)	0.64(0.88)	0.84(0.49)	0.77(0.36)

Table 7: Average pairwise absolute agreement and Cohen's quadratic weighted kappa (in bracket) for annotators for the systems on four aspects described in Section 4.1. BTTS10 indicates the semi-supervised BTTS model with 10% training data.

[4] https://scikit-learn.org/

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 8

□ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1 and Section 8

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

□ B1. Did you cite the creators of artifacts you used?
No response.

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

□ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☑ Did you run computational experiments?

Section 3,4,5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix Section 1

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix Section 1

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 5

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 3.5

D ☑ **Did you use human annotators (e.g., crowdworkers) or research with human participants?**
Section 4

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
In supplementary material

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section 4,1

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
In supplementary material

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.