REFEREE: A REference-FREE Model-Based Metric for Text Simplification

Yichen Huang, Ekaterina Kochmar
Mohamed bin Zayed University of Artificial Intelligence
{yichen.huang, ekaterina.kochmar}@mbzuai.ac.ae

Abstract

Text simplification lacks a universal standard of quality, and annotated reference simplifications are scarce and costly. We propose to alleviate such limitations by introducing REFEREE, a REference-FREE model-based metric with a 3-stage curriculum. REFEREE leverages an arbitrarily scalable pretraining stage and can be applied to any quality standard as long as a small number of human annotations are available. Our experiments show that our metric outperforms existing reference-based metrics in predicting overall ratings and reaches competitive and consistent performance in predicting specific ratings while requiring no reference simplifications at inference time.

Keywords: model-based evaluation, text simplification

1. Introduction

The task of text simplification (TS) aims to reduce the reading and grammatical complexity of text while retaining its meaning and grammaticality (Chandrasekar and Srinivas, 1997). Alongside other text-to-text generation tasks (such as machine translation or summarization), it has been a common practice in TS to use reference-based automatic metrics and evaluate a generated text by comparing it to gold-standard references, typically produced by humans. However, the complexities of TS pose particular challenges for this type of evaluation approaches, and prevalent metrics such as BLEU (Papineni et al., 2002), SARI (Xu et al., 2016) and BERTScore (Zhang et al., 2020) have been shown to correlate poorly with human evaluation (Alva-Manchego et al., 2021; Maddela et al., 2023).

To a large extent, this is due to the absence of a singular, precise definition of what text simplification aims to do and how the output quality should be judged: text simplification may involve lexical, syntactic and conceptual modifications conducted using different operations (e.g. word-swapping, sentence-splitting, and paraphrasing) (Sulem et al., 2018b), and simplification quality is associated with different aspects such as fluency (or grammaticality), meaning preservation (or adequacy) and simplicity. Since reference-based metrics rely on availability of a large enough number of diverse yet high-quality references, the availability of such reference outputs creates a clear bottleneck, as collection of human-produced references is a slow and expensive process. Even though some high quality TS corpora (e.g., Newsela (Xu et al., 2015)) exist, they are still costly to create and often are not open-access (Martin et al., 2018). Finally, there are cases when such references will be impossible to collect: e.g., when there is a need to estimate the quality of text in real time as is increasingly becoming the case for texts generated using large language models (LLMs) (Zhang et al., 2023).

The above reasons motivate development of supervised, model-based evaluation approaches, where a model is trained to mimic human evaluation on given examples, which is applicable to any standard as long as the annotations are available and internally consistent. The need for reliable reference-free evaluation metrics has been expressed before (Specia et al., 2010; Thompson and Post, 2020), and more recently a number of learnable TS metrics have been proposed (Maddela et al., 2023; Zhao et al., 2023; Cripwell et al., 2023). For instance, Maddela et al. (2023) have proposed LENS, where they use RoBERTa-extracted (Liu et al., 2019) representations of (source, simplification, reference) tuples to predict overall quality scores as annotated by humans. Whereas LENS significantly outperforms conventional metrics in correlation with human judgements, it is severely limited by the scarcity of human annotations, with its training data consisting of only 2.4K simplification outputs from 24 systems.

Inspired by the success of BLEURT (Sellam et al., 2020) in machine translation, we propose pretraining on synthesised data and supervision signals as a means to leverage large-scale, unlabeled data and overcome the bottleneck of reference simplifications and human ratings. To facilitate the arbitrarily scalable synthesis of pretraining data, we argue for a reference-free, source-based metric. Specifically, we use existing TS models to produce simplifications for arbitrary source sentences. Given (source, simplification) pairs, we task a model-based metric to predict a range of synthesised supervision signals such as BERTScore, GPT-2 perplexity and model-based simplicity rat-

LREC-COLING 2024, pages 13740–13753

ings. An additional benefit is that, by enabling direct comparison with the source sentence, such a metric can more accurately evaluate criteria such as meaning preservation and relative simplicity (as opposed to source-free metrics such as BLEURT and BERTScore).

In this work, we introduce REFeREE, a model-based metric for text simplification that is reference-free. We propose a curriculum with two pretraining stages and a fine-tuning stage as shown in Figure 1. The first pretraining stage uses reference-free supervision signals and is arbitrarily scalable, allowing us to leverage the large amounts of unlabeled texts. The second pretraining stage relies on both reference-free and reference-based supervision signals. This stage makes use of the readily available TS corpora that do not include human ratings and provides more accurate supervision. Finally, we fine-tune the metric on human ratings such that it is aligned with the specific simplification operations and criteria.

We evaluate our approach on overall ratings from the SimpEval dataset (Maddela et al., 2023) and specific adequacy, fluency and simplicity ratings from the smaller Simplicity-DA (Alva-Manchego et al., 2021) and Human-Likert (Scialom et al., 2021b) datasets. REFeREE correlates better with the overall human ratings, outperforming popular rule-based metrics, BERTScore, BLEURT, LENS and other model-based strong baselines while using a smaller model and requiring less information at inference time. On the smaller Simplicity-DA and Human-Likert datasets, REFeREE overall underperforms Lens but still performs better than conventional metrics and is more consistent across different datasets. Additionally, we perform extensive ablation studies to investigate the effects of each component in our training process.[1]

2. Related Work

The literature on TS models and automatic evaluation is vast. In this section, we provide a brief outline of the types of model-based metrics and then give a more detailed account of the previous work that addresses reference-free evaluation.

2.1. Model-Based Evaluation Metrics

Model-based evaluation metrics have been actively investigated in NLP and particularly machine translation (MT). Metrics fine-tuned on human ratings, such as BLEURT (Sellam et al., 2020; Pu et al., 2021), COMET (Rei et al., 2020), and UniTE (Wan et al., 2022), have been shown to produce good results. In this work, we take direct inspiration from

[1]Code and model checkpoints are available at https://github.com/i-need-sleep/referee.

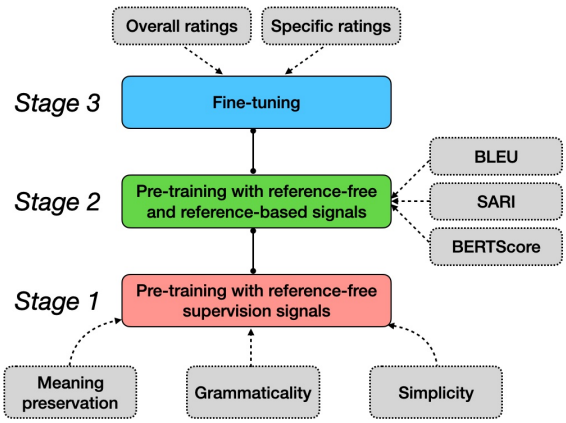

Figure 1: An overview of the proposed curriculum.

BLEURT, a generic metric for natural language generation (NLG). BLEURT produces evaluation from (prediction, reference) pairs and utilizes an arbitrarily scalable pretraining stage where it is trained to predict automatically generated supervision signals (BLEU, BERTScore, etc.) based on synthesized, semantically similar pairs. The pretraining step allows BLEURT to produce good results even when fine-tuned on a reasonably small set of human annotations. However, this method is not immediately applicable to TS evaluation as reference simplifications are difficult to synthesize and TS evaluation involves particular aspects of quality (e.g. simplicity) not considered in BLEURT's pretraining objectives. In response to these challenges, we propose a reference-free, source-based setup with a modified pretraining process.

There also exists a family of unsupervised metrics that do not rely on human ratings and instead compare the embeddings either between the system output and the reference (Zhang et al., 2020) or between the system output and the source sentence (Zhao et al., 2020; Reimers and Gurevych, 2020; Belouadi and Eger, 2023). This line of research is not immediately suitable for our task as text simplification can be evaluated in terms of different aspects, resulting in different ratings even when given the same (source, simplification, reference) tuple.

Another recent trend that addresses automatic evaluation is based on the direct use of Large Language Models (LLMs) for evaluation. Fu et al. (2023) use LLM-predicted conditional probabilities as estimates for NLG quality. Liu et al. (2023) task LLMs to directly produce numeric ratings given descriptions of evaluation criteria through a chain-of-thought (Wei et al., 2022) process. With natural language instruction as an interface for defining evaluation criteria, this type of approach is flexible, data-lean at inference time, and has been shown to

correlate well with human ratings. However, LLMs are compute-hungry and tend to overestimate outputs generated by models similar to themselves (Liu et al., 2023). In addition, most existing LLM-based metrics are aligned with human raters implicitly (through natural language instruction) rather than explicitly (with human ratings). As such, we see this line of work as orthogonal to ours.

2.2. Reference-Free Evaluation Metrics

Traditionally, the outputs of text-to-text generation models have been evaluated using *reference-based metrics* such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), or BERTScore (Zhang et al., 2020). These metrics do not always correlate with human judgments of the generated output quality (Sulem et al., 2018a), and such evaluation paradigm falls short when human references are not available or only a single model is available (Louis and Nenkova, 2013). The quality of the references also has an impact on reference-based evaluation, and, given that evaluation of generated outputs is cognitively demanding and highly subjective (Vasilyev et al., 2020), it is hard to avoid variability across a set of references and gold standard judgements (Harman and Over, 2004).

As a result, development of Quality Estimation (QE) (Specia et al., 2010) and *reference-free metrics* has gained increased attention in the recent years (Louis and Nenkova, 2013; Scialom et al., 2019; Thompson and Post, 2020, inter alia). For summarization, Louis and Nenkova (2013) show that quantifying the similarity between the source text and its summary with appropriately chosen content similarity based measures produces scores which replicate human assessments accurately. Scialom et al. (2019, 2021a) and Vasilyev et al. (2020) evaluate summary quality by measuring how such summaries help with related tasks rather than how well they align with a pre-defined set of references: Scialom et al. (2019, 2021a) introduce metrics based on the intuition that the quality of a generated summary is directly related to the number of relevant questions that can be answered on its basis; and Vasilyev et al. (2020) estimate summary quality measuring the performance boost gained by a pre-trained language model with access to the summary while carrying out its language understanding task on the document's text. In MT, such systems as COMET and its extensions, which use pre-training and subsequent normalization, and Prism (Thompson and Post, 2020), which casts the evaluation task to that of scoring MT output with a zero-shot paraphraser, show results competitive with reference-based models (Rei et al., 2020, 2021); finally, Fonseca et al. (2019) demonstrate that such metrics also highly correlate with human judgments while alleviating the need for references, thus suggesting that reference-free evaluation is a promising direction for future research.

2.3. Reference-Free TS Evaluation

In TS, researchers have also been investigating application of reference-free measures (Temnikova and Maneva, 2013; Kajiwara and Fujita, 2017, inter alia): for instance, Temnikova and Maneva (2013) propose to evaluate TS quality extrinsically via human reading comprehension. Other early works on automatic reference-free evaluation rely on feature-based classification approaches aimed at specific aspects of simplification (e.g., lexical, syntactic, structural). For instance, Štajner et al. (2014) investigate applicability of popular MT evaluation metrics to the TS systems outputs and the corresponding original sentences, and demonstrate their potential in replacing human assessment of TS systems aimed at syntactic simplifications and content reduction. Kajiwara and Fujita (2017) show that a classification model utilizing alignment-based semantic features is capable of reliably predicting TS system quality when lexical simplifications are also involved as is the case with the QATS 2016 dataset (Štajner et al., 2016). Experiments of Martin et al. (2018) on the same data demonstrate that n-gram-based MT metrics correlate the most with human judgment of grammaticality and meaning preservation, whereas simplicity is best evaluated by basic length-based metrics. Finally, Sulem et al. (2018b) focus on the structural aspects of TS and propose SAMSA, a system that uses decomposition of the input based on its semantic structure and compares it to the output.

A different line of research investigates the use of large-scale pre-trained models for direct quality estimation in TS systems: a notable example is Kriz et al. (2020), who propose Simple-QE, a BERT-based QE model adapted from prior summarization work, and show that it correlates well with human quality judgments. Scialom et al. (2021b) present a BERTScore-based (Zhang et al., 2020) adaptation of their QuestEval metric (Scialom et al., 2021a) for TS and show that it yields competitive results along the meaning preservation dimension, with considerable improvement over BLEU and SARI. At the same time, their analysis suggests that datasets commonly considered in TS research show a considerable level of spurious correlations between different dimensions, with fluency being highly correlated with meaning preservation and simplicity. To that end, they release Human-Likert, a new large corpus of human evaluations devoid of such spurious correlations.

Recently, Zhao et al. (2023) have proposed BETS, a reference-free TS metric aggregating a simplicity score and an adequacy score. The simplicity branch is trained on pairs of complex and

Training stage	Data	Supervision signals
Pretraining (Stage 1)	OpenWebText	**Meaning preservation**: SBERT, self-BLEU, self-BERTScore **Fluency**: Source perplexity, simplification perplexity **Simplicity**: Source FKGL, simplification FKGL, source simplicity, simplification simplicity
Pretraining (Stage 2)	Newsela (test) WikiSmall (test) WikiLarge (test)	All stage 1 objectives + BLEU, SARI, BERTScore
Fine-tuning (Stage 3)	**Overall**: SIMPEVAL **Specific**: SIMPLICITY-DA, HUMAN-LIKERT	Human ratings

Table 1: An overview of the data and supervision signals used in each training stage. For the final stage, we fine-tune and evaluate REFEREE separately for each dataset and quality aspect.

simple phrases, and the adequacy branch is based on word embedding similarity akin to BERTScore. At the same time, Cripwell et al. (2023) propose to evaluate TS quality with SLE, a reference-free simplicity metric trained on softened reading levels from Newsela (Xu et al., 2015). These metrics focus on specific simplification qualities and evaluation criteria: BETS primarily considers lexical simplification, with its simplicity branch unable to return a score when all words in the simplified sentence are present in the original sentence (as would be the case with simplification by deletion or splitting), and SLE is not trained nor tested on machine simplifications with adequacy or fluency issues.

By contrast, we propose a metric that is compatible with any simplification operation as it is fine-tuned end-to-end on human ratings. To the best of our knowledge, the only learnable model-based TS metric aligned with human ratings is LENS (Maddela et al., 2023), which adopts a COMET-like approach and relies on references.

3. Methodology

REFEREE is based on a pretrained `DeBERTa-v3-base` model (He et al., 2023) and takes as input delimited pairs of source sentences and system outputs. The DeBERTa-extracted sequence embedding is passed into a linear regression head for each supervision signal. The training process consists of three stages: (1) an arbitrarily scalable pretraining stage with reference-free supervision signals, (2) a second pretraining stage with both reference-free and reference-based supervision signals, and (3) a fine-tuning stage with human ratings. An overview of the data and supervision signals is shown in Table 1. We describe the three stages in the following subsections. More implementation details are reported in Appendix A.

3.1. Pretraining with Reference-Free Supervision Signals

The first pretraining stage aims to learn important aspects of simplification quality directly from text. It involves a collection of reference-free supervision signals and is designed to be arbitrarily scalable such that the metric can leverage large amounts of unlabeled text. Based on the common criteria of text simplification quality (Martin et al., 2018; Kriz et al., 2020; Scialom et al., 2021b, inter alia), we select a range of supervision signals measuring meaning preservation (also referred as adequacy), fluency (or grammaticality) and simplicity:

- **Meaning preservation (adequacy)** focuses on *how well the TS output preserves the meaning of the original text*. For this, we use the cosine embedding distance from SBERT embeddings (Reimers and Gurevych, 2019) as well as self-BLEU and self-BERTScore measured against the source sentence. In this way, our metric can capture both lexical overlaps and paraphrases.
- **Fluency (grammaticality)** aims to measure *how well-formed the TS output is*. For this, we include the perplexity for both the source sentence and the system output as measured by GPT-2 (Radford et al., 2019). Our intuition is that if the content of the source and the system output is similar, then the difference between their perplexities reflects the difference in fluency.
- **Simplicity** focuses on *the extent to which the output is easier to read and understand than the original text*. Here, we utilize the FKGL score (Kincaid et al., 1975) and a model-based readability score for both the source and the system output. We use an ALBERT-based (Lan et al., 2020) system trained on the CommonLit dataset.[2]

To generate pairs of source sentences and machine simplifications, we use a small subset of the OpenWebText dataset (Gokaslan and Cohen, 2019) of approximately 200K sentences. To obtain a range of good and bad simplifications, we use outputs from high-performing models and augment the dataset with degraded simplifications. We use

[2] https://github.com/mathislucka/kaggle_clrp_1st_place_solution

outputs from MUSS (Martin et al., 2022) as well as 5-shot results from GPT-3.5-turbo (OpenAI, 2023) and GPT-3-Curie (Brown et al., 2020). Based on the results of preliminary experiments, we augment 40% of the outputs by random deletion, scrambling and swapping of the original and simplified sentences. As no annotated reference simplification is required, this stage is arbitrarily scalable, and the set of supervision signals is extendable.

3.2. Pretraining with Reference-Free and Reference-Based Supervision Signals

As highlighted earlier, TS data with human ratings are scarce, which creates a bottleneck for TS evaluation. However, there exist several TS corpora of aligned complex and human-simplified sentences, and in the second stage, we utilize this data in the form of reference-based supervision signals to provide more accurate supervision. Specifically, the second pretraining stage includes BLEU, SARI and BERTScore as supervision signals in addition to the reference-free signals from Stage 1. We utilize the Newsela (Xu et al., 2015), WikiSmall and WikiLarge (Zhang and Lapata, 2017) test sets, totalling at 1,536 (source, reference) pairs. In addition to the outputs from the models used in the previous stage, we include outputs from EditNTS (Dong et al., 2019), DRESS (Zhang and Lapata, 2017), Hybrid (Narayan and Gardent, 2014), and PBMT-R (Wubben et al., 2012). Due to the reliance on reference simplifications, this stage is scalable only in terms of the number of simplification systems.

3.3. Fine-tuning

Finally, we fine-tune the metric on human ratings such that it is aligned with the particular criteria used in each dataset. We consider fine-tuning on overall as well as specific ratings. For overall ratings, we use the SimpEval (Maddela et al., 2023) corpus which contains overall quality ratings from five annotators. Following Maddela et al. (2023), we use the $\text{SimpEval}_{\text{PAST}}$ and SimpEval_{2022} subsets of the SimpEval corpus respectively for training and evaluation. $\text{SimpEval}_{\text{PAST}}$ contains ratings for 2.4K system outputs from 24 systems and is based on a subset of the TurkCorpus (Xu et al., 2016). SimpEval_{2022} is designed to present a more challenging scenario and contains 360 simplifications from 6 systems for a new, curated set of sentences that are longer and discuss recent events. In particular, SimpEval_{2022} includes only higher-quality simplifications from GPT-3.5 (OpenAI, 2023), T5 (Raffel et al., 2020), MUSS (Martin et al., 2022) and human annotators. We report the aggregated results from three runs.

To evaluate our method on learning specific scores, we utilize the Simplicity-DA (Alva-Manchego et al., 2021) and Human-Likert (Scialom et al., 2021) datasets. Simplicity-DA contains 600 system outputs from six systems from the TurkCorpus (Xu et al., 2016) test set, each with annotations on adequacy, fluency and simplicity. Human-Likert follows the same format and contains 112 human-written simplifications from the ASSET(Alva-Manchego et al., 2020) and TurkCorpus test sets. The smaller size of the two datasets presents a challenge to learnable metrics. We fine-tune and evaluate REFeREE separately on the different aspects. We split the datasets into training, validation and test sets by source sentences with a 4-1-1 ratio and report the averaged results on five runs with non-overlapping test sets.

4. Experimental Results

We fine-tune and evaluate REFeREE separately on the three datasets. For each stage, the model is trained on the unweighted average of L2 losses from the training signals. We use the Adam optimizer (Kingma and Ba, 2015) with $\epsilon = 10^{-6}$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$, learning rates of 10^{-5}, 10^{-5}, and 10^{-7} for the three stages, and perform early-stopping based on the development set performance.

Human ratings are aggregated by taking the average. For Simplicity-DA and Human-Likert, we report the Pearson correlation r. For SimpEval, as the ratings for different simplification operations (deletion, paraphrase and splitting) are separately collected and not comparable, we follow Maddela et al. (2023) and report the Kendall Tau-like coefficient τ (Bojar et al., 2017), which has a range between -1 and 1 and is defined as:

$$\tau = \frac{|Concordant| - |Discordant|}{|Concordant| + |Discordant|}$$

where $Concordant$ is the set of pair-wise rankings where the metric agrees with human ratings on simplifications for the same source sentence, and $Discordant$ is the set of rankings where they disagree.

We compare our metric with non-learnable metrics (FKGL, BLEU, SARI, self-BERTScore and BERTScore), as well as BLEURT (Pu et al., 2021), Lens, BETS, and SLE.[3] For a fair comparison with Lens, we also include a variant of REFeREE with

[3]We include the precision score of BERTScore in addition to the F1 score as Alva-Manchego et al. (2021) observe that it correlates better with human judgments. BLEURT was originally fine-tuned on WMT datasets (Barrault et al., 2019). We also experiment with fine-tuning it on the simplification datasets. We compare against the

Metric	$\tau_{paraphrase}$ ↑	$\tau_{splitting}$ ↑	τ_{all} ↑
FKGL	-0.556	-0.310	-0.356
BLEU	-0.048	-0.054	-0.033
SARI	0.397	0.264	0.289
BERTScore$_{\text{F1}}$	0.175	0.023	0.052
BERTScore$_{\text{precision}}$	0.238	0.093	0.112
self-BERTScore$_{\text{F1}}$	-0.174	-0.333	-0.300
self-BERTScore$_{\text{precision}}$	-0.079	-0.348	-0.300
BLEURT$_{\text{WMT}}$	0.055	0.073	0.030
BLEURT$_{\text{fine-tuned}}$	0.270 ± 0.113	0.132 ± 0.023	0.163 ± 0.040
BETS	-0.302	-0.349	-0.331
SLE	0.492	0.256	0.295
LENS$_{k=3}$	0.429	0.333	0.331
REFEREE	0.481 ± 0.015	$\mathbf{0.341 \pm 0.029}$	0.360 ± 0.020
REFEREE (RoBERTa)	$\mathbf{0.534 \pm 0.030}$	0.328 ± 0.019	$\mathbf{0.368 \pm 0.018}$

Table 2: Results on the SIMPEVAL$_{2022}$ dataset for different operation types. We use the official checkpoints for off-the-shelf metrics. For the metrics we trained (REFEREE and BLEURT$_{\text{fine-tuned}}$), we report the aggregated results from three runs. We follow Maddela et al. (2023) and report the Kendall Tau-like coefficient on filtered pairs where all three annotators agree with the ranking order and the unnormalized score difference (out of 100) is larger than five for at least two annotators. Since each operation group is rated separately for the dataset, we do not compare simplifications of different operation types. For the same reason, the overall Pearson correlation is not compatible with this dataset.

a RoBERTa-large backbone. Finally, to investigate the contribution of the three training stages, data augmentation and training signals, we also report ablation results on SIMPEVAL.

4.1. Overall Ratings

Results on the SIMPEVAL$_{2022}$ dataset are shown in Table 2.[4] Non-learnable metrics perform poorly, with FKGL, BLEU and self-BERTScore having more discordant pairs than concordant ones, demonstrating the need for learnable metrics. In particular, self-BERTScore measures sentence similarity and may mistakenly punish simplifications that remove non-essential information or inadequately punish under-simplifications. The off-the-shelf and fine-tuned BLEURT metrics also perform poorly as the BLEURT pretraining process mainly considers semantic similarity and does not adequately encompass all aspects of simplification quality. Somewhat surprisingly, BETS performs very poorly for all types of simplification. We hypothesize that this is because it is primary designed for lexical simplification and cannot effectively evaluate simplification by syntactic changes. SLE, a metric trained to predict simplicity, performs relatively well on this dataset, suggesting that the overall quality ratings are correlated with sentence simplicity and that the simplifications in SIMPEVAL$_{2022}$ have relatively few adequacy and fluency issues.

REFEREE outperforms LENS, which uses a larger model (354M parameters compared with 226M parameters) and relies on more information at inference time. Changing to the larger RoBERTa-large backbone results in slightly improved overall performance, with a significant improvement in evaluating paraphrases. A further inspection reveals that compared with LENS, REFEREE is more effective at evaluating machine simplifications and relatively underperforms when handling human simplifications. Specifically, REFEREE results in a KendallTau-like coefficient of 0.310 when evaluating sentence pairs involving human simplifications and 0.500 when evaluating other pairs whereas the results for LENS are 0.371 and 0.273, respectively. We suspect that this is because REFEREE is less exposed to human-written simplifications during pretraining compared with machine simplifications.

4.2. Specific Ratings

Tables 3 and 4 show the results on the SIMPLICITY-DA and HUMAN-LIKERT datasets for adequacy, fluency and simplicity. Overall, REFEREE underperforms LENS but still performs better than conventional metrics. On SIMPLICITY-DA, we observe that most source-based metrics (self-BERTScore, LENS and REFEREE) outperform source-free metrics, suggesting that having direct access to the source sentence helps with measuring meaning preservation. Self-BERTScore, with the highest perfor-

highest-performing LENS variant trained using the top-3 references and the highest-performing SLE variant measuring relative simplicity and trained on softened labels. In the cases where the simplicity branch of BETS fails to return a score, we assign a simplicity score of 0.

[4]We observe slightly different results than those reported in Maddela et al. (2023), which is likely due to the use of different implementations of the evaluation metrics.

Metric	Adequacy $r\uparrow$	Fluency $r\uparrow$	Simplicity $r\uparrow$
FKGL	0.064 ± 0.164	0.083 ± 0.207	0.099 ± 0.117
BLEU	0.354 ± 0.153	0.317 ± 0.144	0.221 ± 0.145
SARI	0.258 ± 0.066	0.164 ± 0.077	0.180 ± 0.094
BERTScore$_{\text{F1}}$	0.569 ± 0.075	0.462 ± 0.108	0.362 ± 0.066
BERTScore$_{\text{Precision}}$	0.513 ± 0.095	0.480 ± 0.111	0.426 ± 0.074
self-BERTScore$_{\text{F1}}$	$\mathbf{0.727 \pm 0.044}$	0.528 ± 0.083	0.390 ± 0.045
self-BERTScore$_{\text{Precision}}$	0.687 ± 0.069	0.566 ± 0.054	0.481 ± 0.041
BLEURT$_{\text{WMT}}$	0.595 ± 0.082	0.437 ± 0.172	0.323 ± 0.111
BLEURT$_{\text{fine-tuned}}$	0.096 ± 0.229	0.384 ± 0.158	0.150 ± 0.228
BETS	0.592 ± 0.050	0.367 ± 0.094	0.155 ± 0.065
SLE	-0.329 ± 0.097	-0.128 ± 0.155	0.018 ± 0.082
LENS$_{k=3}$	0.636 ± 0.069	$\mathbf{0.758 \pm 0.059}$	$\mathbf{0.732 \pm 0.094}$
REFEREE	0.622 ± 0.079	0.478 ± 0.045	0.366 ± 0.126
REFEREE (RoBERTa)	0.633 ± 0.038	0.483 ± 0.57	0.427 ± 0.058

Table 3: Results on the SIMPLICITY-DA dataset. The LENS model is trained on SIMPEVAL$_{\text{PAST}}$ and not fine-tuned on this dataset. The dataset is not compatible with the Kendall Tau-like coefficient as it mostly does not contain simplifications from different systems for the same source sentence.

Metric	Adequacy $r\uparrow$	Fluency $r\uparrow$	Simplicity $r\uparrow$
FKGL	0.111 ± 0.125	-0.169 ± 0.136	-0.385 ± 0.226
BLEU	0.280 ± 0.163	0.316 ± 0.183	0.157 ± 0.209
SARI	0.139 ± 0.120	0.236 ± 0.089	0.445 ± 0.108
BERTScore$_{\text{F1}}$	0.280 ± 0.150	0.214 ± 0.042	0.105 ± 0.130
BERTScore$_{\text{Precision}}$	0.266 ± 0.192	0.433 ± 0.065	0.321 ± 0.077
self-BERTScore$_{\text{F1}}$	0.421 ± 0.169	0.016 ± 0.103	-0.385 ± 0.108
self-BERTScore$_{\text{Precision}}$	0.345 ± 0.210	0.175 ± 0.107	-0.255 ± 0.128
BLEURT$_{\text{WMT}}$	0.441 ± 0.047	0.247 ± 0.022	0.093 ± 0.072
BLEURT$_{\text{fine-tuned}}$	$\mathbf{0.472 \pm 0.110}$	0.251 ± 0.150	0.077 ± 0.109
BETS	0.375 ± 0.167	-0.114 ± 0.106	-0.513 ± 0.066
SLE	-0.206 ± 0.142	0.186 ± 0.099	0.532 ± 0.072
LENS$_{k=3}$	0.201 ± 0.122	$\mathbf{0.561 \pm 0.057}$	$\mathbf{0.561 \pm 0.055}$
REFEREE	0.425 ± 0.127	0.308 ± 0.108	0.322 ± 0.075
REFEREE (RoBERTa)	0.386 ± 0.120	0.292 ± 0.171	0.528 ± 0.083

Table 4: Results on the HUMAN-LIKERT dataset. The LENS model is trained on SIMPEVAL$_{\text{PAST}}$ and not fine-tuned on this dataset. The dataset is not compatible with the Kendall Tau-like coefficient as it mostly does not contain simplifications from different systems for the same source sentence.

mance on adequacy, also performs well in measuring fluency and simplicity. This is likely due to the high intra-correlation between the three aspects in Simplicity-DA (e.g. inadequate simplifications are likely not fluent and difficult to understand) as pointed out by (Scialom et al., 2021b). SLE, despite performing well on SIMPLICITY-DA when controlling for adequacy and fluency (Cripwell et al., 2023), performs poorly on the unfiltered dataset as it is not exposed to the lower-quality machine simplifications during its training process. The relative underperformance of REFEREE is likely because the dataset includes outputs from several dated systems absent in our pretraining process. Incorporating more systems and refining the augmentation process will likely lead to improvements.

We observe that the performance of most metrics is inconsistent between SIMPLICITY-DA and HUMAN-LIKERT. For instance, LENS performs much worse in predicting adequacy scores on HUMAN-LIKERT than on SIMPLICITY-DA. Aside from HUMAN-LIKERT having lower inter-correlations, this can also be due to the variances in the collection of human ratings (e.g. annotators and criteria). Despite having lower overall performance, REFEREE performs more consistently between the datasets as it can be fine-tuned to be aligned with the particularities of each set of ratings.

4.3. Ablation Experiments

Finally, to shed light on how the different components of REFEREE affect the overall performance, we report the ablation study results on SIMPEVAL$_{2022}$ with respect to the training regime, the types of supervision signals and data augmenta-

Metric	$\tau_{paraphrase} \uparrow$	$\tau_{splitting} \uparrow$	$\tau_{all} \uparrow$
REFEREE	$\mathbf{0.481 \pm 0.015}$	0.341 ± 0.029	0.360 ± 0.020
Stage 1 + fine-tuning	0.439 ± 0.054	0.370 ± 0.041	$\mathbf{0.374 \pm 0.031}$
Stage 2 + fine-tuning	0.365 ± 0.052	0.240 ± 0.058	0.254 ± 0.058
Fine-tuning only	0.354 ± 0.030	0.220 ± 0.045	0.242 ± 0.037
REFEREE, all stages			
w/o fluency	0.376 ± 0.015	0.328 ± 0.007	0.335 ± 0.003
w/o meaning	0.407 ± 0.030	$\mathbf{0.372 \pm 0.013}$	0.358 ± 0.006
w/o simplicity	0.418 ± 0.015	0.305 ± 0.015	0.309 ± 0.016
w/o augmentation	0.439 ± 0.015	0.271 ± 0.006	0.289 ± 0.009

Table 5: Ablation results on SIMPEVAL$_{2022}$.

tion. Comparing variants of REFEREE with different training regimes, we find that the first pretraining stage significantly improves the performance thanks to its relatively large scale, leading to an increase of over 0.13 in the Kendall Tau-like coefficient for all operation types compared with fine-tuning only. The second pretraining stage appears limitedly helpful when compared with fine-tuning only and even results in slightly degraded performance when combined with the first pretraining stage. This can be due to potentially varying quality of the references and the small dataset size for this stage, which may more easily lead to overfitting. This further signifies the utility of the arbitrarily scalable first pretraining stage as a means to improve model performance under the scarcity of human-annotated ratings.

We also experiment with ablating types of supervision signals for meaning preservation (BLEU, self-BLEU, BERTScore and self-BERTScore), fluency (GPT-2 perplexity) and simplicity (FKGL and CommonLit readability). We observe that the supervision signals for meaning preservation only slightly influence the performance, which is likely because the machine simplifications in SimpEval2022 are of high quality and generally preserve the meaning of the complex sentences. The signals for fluency and simplicity seem to play a more impactful role, likely because the systems in SimpEval2022 (GPT-3.5, T5, MUSS and humans) produce simplifications in different styles and of varying fluency. This explains why our metric outperforms non-specialized metrics such as BERTScore which primarily focus on meaning preservation.

Ablating data augmentation leads to a significant decrease in the metric's performance. This is because the first pre-training stage involves relatively high-quality simplification produced by MUSS and GPT, and data augmentation is an effective way of generating lower-quality simplifications that complement these system outputs.

5. Discussion

Despite promising results demonstrated in this work, we also recognize that there are limitations of the proposed method. In this section, we discuss such limitations, outlining potential directions for future research.

First of all, due to the data limitations, REFEREE is only fine-tuned and evaluated on sentence-level simplifications in the English news and Wikipedia domains. Its performance on other languages, domains and simplification setups (e.g. document-level simplification (Sun et al., 2021) and elaborative simplification (Srikanth and Li, 2021)) awaits further investigation. Whereas our reference-free pretraining stage is arbitrarily scalable in design, we only experimented with a reasonably small pretraining dataset of around 200K source sentences simplified by three systems. Further experiments are required to determine how the performance of the metric scales with the dataset size and the number of simplification systems. In addition, since the metric is fine-tuned on small datasets, its out-of-domain performance on other datasets and simplification systems is not guaranteed.

Finally, despite the increasing need for reference-free evaluation metrics and the development of multiple reference-free approaches, the applicability of such metrics needs to be carefully considered. Deutsch et al. (2022) highlight that the metrics can be over-optimized at test time, and may be biased both towards models similar to their backbones and against higher-quality outputs produced by humans. We agree with Deutsch et al. (2022) that reference-free metrics should be used as diagnostic tools and with Louis and Nenkova (2013) that these metrics should complement high-quality human evaluation. However, we also contend that this still makes them useful during the rapid prototyping of new systems where human evaluations are costly, difficult or sometimes impossible to collect.

6. Conclusion

We propose REFEREE, a reference-free model-based metric for text simplification with a 3-stage

curriculum, including an arbitrarily scalable pretraining using reference-free supervision signals as well as pretraining with both reference-free and reference-based supervision signals, and a finetuning stage with human ratings. Our experiments show that our metric is effective and flexible, attaining competitive performance in evaluating with respect to both general and specific ratings of the quality of the text simplification system outputs.

Since the formulation of our metric is largely generalizable, it can be modified and applied to other conditional natural language generation tasks such as abstractive summarization, among others. This calls for an investigation into task-agnostic and multi-task supervision, which we leave as future work.

Acknowledgements

This work is financially supported by Mohamed bin Zayed University of Artificial Intelligence (MBZUAI) and is supported by the Campus Super Computing Center at MBZUAI. We thank the anonymous reviewers for their valuable feedback.

7. Bibliographical References

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.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 Conference on Machine Translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Jonas Belouadi and Steffen Eger. 2023. UScore: An Effective Approach to Fully Unsupervised Evaluation Metrics for Machine Translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 358–374, Dubrovnik, Croatia. Association for Computational Linguistics.

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017. Results of the WMT17 Metrics Shared Task. In *Proceedings of the Second Conference on Machine Translation*, pages 489–513, Copenhagen, Denmark. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Raman Chandrasekar and Bangalore Srinivas. 1997. Automatic induction of rules for text simplification. *Knowledge-Based Systems*, 10(3):183–190.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2023. Simplicity Level Estimate (SLE): A Learned Reference-Less Metric for Sentence Simplification. *arXiv preprint arXiv:2310.08170*.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2022. On the Limitations of Reference-Free Evaluations of Generated Text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10960–10977, Abu Dhabi, United Arab Emirates. 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.

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.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire.

Donna Harman and Paul Over. 2004. The Effects of Human Variation in DUC Summarization Evaluation. In *Text Summarization Branches Out*, pages 10–17, Barcelona, Spain. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tomoyuki Kajiwara and Atsushi Fujita. 2017. Semantic Features Based on Word Alignments for Estimating Quality of Text Simplification. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 109–115, Taipei, Taiwan. Asian Federation of Natural Language Processing.

J. Peter Kincaid, Robert P. Fishburne, 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.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR (Poster)*.

Reno Kriz, Marianna Apidianaki, and Chris Callison-Burch. 2020. Simple-QE: Better Automatic Quality Estimation for Text Simplification. *arXiv preprint arXiv:2012.12382*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *International Conference on Learning Representations*.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach.

Annie Louis and Ani Nenkova. 2013. Automatically assessing machine summary content without a gold standard. *Computational Linguistics*, 39(2):267–300.

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.

Jonathan Mallinson, Aliaksei Severyn, Eric Malmi, and Guillermo Garrido. 2020. FELIX: Flexible Text Editing Through Tagging and Insertion. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1244–1255, Online. Association for Computational Linguistics.

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.

OpenAI. 2023. ChatGPT: Optimizing language models for dialogue.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Amy Pu, Hyung Won Chung, Ankur Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning Compact Metrics for MT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 751–762, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2021. An Exploratory Analysis of Multilingual Word-Level Quality Estimation with Cross-Lingual Transformers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 434–440, Online. Association for Computational Linguistics.

Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are References Really Needed? Unbabel-IST 2021 Submission for the Metrics Shared Task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. 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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021a. QuestEval: Summarization Asks for Fact-based Evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers Unite! Unsupervised Metrics for Reinforced Summarization Models. 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 3246–3256, Hong Kong, China. Association for Computational Linguistics.

Thomas Scialom, Louis Martin, Jacopo Staiano, Eric Villemonte de La Clergerie, and Benoît Sagot. 2021b. Rethinking automatic evaluation in sentence simplification. *arXiv preprint arXiv:2104.07560*.

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.

Lucia Specia, Dhwaj Raj, and Marco Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine translation*, 24:39–50.

Neha Srikanth and Junyi Jessy Li. 2021. Elaborative Simplification: Content Addition and Explanation Generation in Text Simplification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5123–5137, Online. Association for Computational Linguistics.

Sanja Štajner, Ruslan Mitkov, and Horacio Saggion. 2014. One Step Closer to Automatic Evaluation of Text Simplification Systems. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 1–10, Gothenburg, Sweden. Association for Computational Linguistics.

Sanja Štajner, Maja Popovic, Horacio Saggion, Lucia Specia, and Mark Fishel. 2016. Shared task on quality assessment for text simplification. *Training*, 218(95):192.

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.

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

Irina Temnikova and Galina Maneva. 2013. The C-score – proposing a reading comprehension metrics as a common evaluation measure for text simplification. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 20–29, Sofia, Bulgaria. Association for Computational Linguistics.

Brian Thompson and Matt Post. 2020. Automatic Machine Translation Evaluation in Many Languages via Zero-Shot Paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

Oleg Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020. Fill in the BLANC: Human-free quality estimation of document summaries. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 11–20, Online. Association for Computational Linguistics.

Yu Wan, Dayiheng Liu, Baosong Yang, Haibo Zhang, Boxing Chen, Derek Wong, and Lidia Chao. 2022. UniTE: Unified Translation Evaluation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8117–8127, Dublin, Ireland. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

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.

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 595–605. Association for Computational Linguistics.

Yue Zhang, Leyang Cui, Enbo Zhao, Wei Bi, and Shuming Shi. 2023. RobustGEC: Robust Grammatical Error Correction Against Subtle Context Perturbation.

Wei Zhao, Goran Glavaš, Maxime Peyrard, Yang Gao, Robert West, and Steffen Eger. 2020. On the Limitations of Cross-lingual Encoders as Exposed by Reference-Free Machine Translation Evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1656–1671, Online. Association for Computational Linguistics.

Xinran Zhao, Esin Durmus, and Dit-Yan Yeung. 2023. Towards reference-free text simplification

evaluation with a BERT Siamese network architecture. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13250–13264, Toronto, Canada. Association for Computational Linguistics.

8. Language Resource References

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, Carolina Scarton, and Lucia Specia. 2021. The (Un)Suitability of Automatic Evaluation Metrics for Text Simplification. *Computational Linguistics*, 47(4):861–889.

Aaron Gokaslan and Vanya Cohen. 2019. OpenWebText Corpus.

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, Eric Villemonte de La Clergerie, and Benoît Sagot. 2021. Rethinking automatic evaluation in sentence simplification. *arXiv preprint arXiv: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.

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.

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 595–605. Association for Computational Linguistics.

A. Implementation Details

A.1. Data

For the first pretraining stage, we use source sentences from the first 10 volumes of the OpenWebText dataset. We use the `muss_en_wikilarge_mined` checkpoint for the MUSS model (Martin et al., 2022). For the GPT models, we randomly sample five source-simplification pairs from the TurkCorpus dataset (Xu et al., 2016) as in-context examples. We use the following prompt template:

- System prompt: *You are a helpful assistant that simplifies English sentences, making them easier to read while preserving key meanings.*
- User prompt: *Follow the examples and simplify the sentence, making it easier to read while preserving key meanings. Reply with only the simplified sentence. Sentence:* {...} *Simplification:* {...} *... Sentence:* {...}

with the system prompt only applicable to the `GPT-3.5-turbo` model. For the models in the second pretraining stage, we use the model outputs as published by their authors.

For the pretraining stages, we randomly select 40% of the system outputs for augmentation. Each instance selected for augmentation is assigned to be augmented by deletion, scrambling or by swapping the complex and simplified sentences respectively with probabilities of 0.3, 0.3 and 0.4. We use the `NLTK Tree Bank Word Tokenizer`[5] and uniformly randomly select one to four words for deletion and one to five words for scrambling.

We use the HuggingFace `Evaluate` implementation[6] of BLEU, the Sentence Transformer implementation of `all-distillroberta-v1`[7] for SBERT, the HuggingFace `Evaluate` implementation[8] of GPT-2 perplexity and the EASSE (Alva-Manchego et al., 2019) implementation of SARI. For FKGL, we use the syllable and lexicon counts calculated using the `Textstat` package.[9] Each supervision signal is normalized across the dataset.

[5] https://www.nltk.org/api/nltk.tokenize.TreebankWordTokenizer.html

[6] https://huggingface.co/spaces/evaluate-metric/bleu

[7] sentence-transformers/all-distilroberta-v1

[8] https://huggingface.co/spaces/evaluate-metric/perplexity

[9] https://pypi.org/project/textstat/

A.2. Model Implementation and Training

We base REFEREE on the HuggingFace implementation [10] of the `DeBERTa-v3-base` model with 12 layers and a hidden size of 768. The model takes as input the source and simplified sentences delimited by a `<SEP>` token. We use the embedding corresponding to the `<BOS>` token as the sequence embedding and feed it into separate linear regression heads for each supervision signal.

For each stage, the model is trained on the unweighted sum of L2 losses from the training signals. We use the Adam optimizer (Kingma and Ba, 2015) with $\epsilon = 10^{-6}$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. For the three stages, we respectively train the model for a maximum of three, 30, and 50 epochs with a learning rate of 10^{-5}, 10^{-5}, and 10^{-7} and apply early-stopping based on the loss on development sets.

For the fine-tuned BLEURT model, we start from the `BLEURT-20-D12` checkpoint[11] fine-tuned on WMT and reinitialize the regression head. The fine-tuning follows the same process as REFEREE.

[10] https://huggingface.co/microsoft/deberta-v3-base

[11] https://github.com/lucadiliello/bleurt-pytorch