LENS: A Learnable Evaluation Metric for Text Simplification

Mounica Maddela[*], **Yao Dou**[*], **David Heineman, Wei Xu**
School of Interactive Computing
Georgia Institute of Technology
{mmaddela3, douy, david.heineman}@gatech.edu; wei.xu@cc.gatech.edu

http://lens-score.com/

Abstract

Training learnable metrics using modern language models has recently emerged as a promising method for the automatic evaluation of machine translation. However, existing human evaluation datasets for text simplification have limited annotations that are based on unitary or outdated models, making them unsuitable for this approach. To address these issues, we introduce the SIMPEVAL corpus that contains: SIMPEVAL$_{PAST}$, comprising 12K human ratings on 2.4K simplifications of 24 past systems, and SIMPEVAL$_{2022}$, a challenging simplification benchmark consisting of over 1K human ratings of 360 simplifications including GPT-3.5 generated text. Training on SIMPEVAL, we present LENS, a **L**earnable **E**valuatio**n** Metric for Text **S**implification. Extensive empirical results show that LENS correlates much better with human judgment than existing metrics, paving the way for future progress in the evaluation of text simplification. We also introduce RANK & RATE, a human evaluation framework that rates simplifications from several models in a list-wise manner using an interactive interface, which ensures both consistency and accuracy in the evaluation process and is used to create the SIMPEVAL datasets.

1 Introduction

Text simplification is a text-to-text generation task that aims to make a text easier to read while preserving its original meaning (Saggion, 2017). Automatic evaluation of text simplification is challenging because a sentence can be simplified in many ways, such as paraphrasing complex words, deleting insignificant information, and splitting long sentences into shorter ones. An ideal automatic metric should accommodate these diverse choices while capturing semantic similarity and fluency. However, existing metrics such as SARI (Xu et al., 2016) and BERTScore (Zhang et al., 2020) struggle

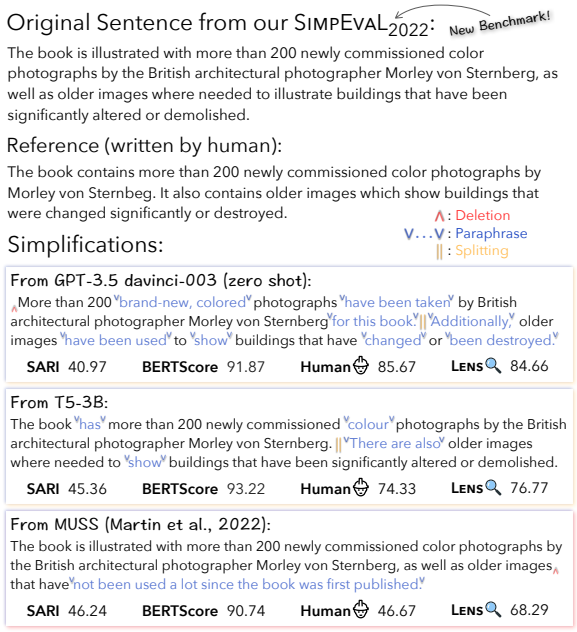


Figure 1: Automatic metric and human evaluation scores on simplifications of a complex sentence from our SIMPEVAL$_{2022}$. LENS achieves the best correlation with humans. SARI penalizes simplifications that deviate from the reference (first two examples), and BERTScore fails to penalize hallucinations with high lexical overlap (third example).

to capture all the aspects and achieve a high correlation with human evaluation (Alva-Manchego et al., 2021) (see Figure 1). These metrics fail even more when evaluating high-quality systems that have close performance, calling for a more robust and accurate metric for text simplification.

Prior work on machine translation (Rei et al., 2020; Sellam et al., 2020) has seen the success of using language models as an automatic metric by training them on human judgments, such as Direct Assessment (DA) that rates generations on a 0-100 scale. However, existing human evaluation datasets (Alva-Manchego et al., 2021; Sulem et al., 2018) for text simplification are not suitable for training metrics because they include a limited number of annotations or systems. Besides, these datasets do not include state-of-the-art generation models such

[*]Equal contribution.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 16383–16408

as GPT-3.5 (Ouyang et al., 2022) that have been shown to generate human-like quality text (Dou et al., 2022; Goyal et al., 2022).

In this work, we introduce LENS, a **L**earnable **E**valuatio**n** Metric for Text **S**implification. LENS is the first supervised metric for the task and uses an adaptive ranking loss to promote fair comparison of simplifications that have undergone different edits (i.e., splitting, paraphrasing, and deletion). To train LENS, we collect 12K human judgments on 2.4K simplifications by 24 simplification systems from the literature, which we name as the SIMPEVAL$_{\text{PAST}}$ dataset. We also create SIMPEVAL$_{2022}$ to evaluate LENS and other metrics in a realistic and challenging setting of assessing simplifications by state-of-the-art language models on the more recent, complex, and longer sentences published on Wikipedia after Oct 22nd, 2022. SIMPEVAL$_{2022}$ contains over 1K human ratings on 360 simplifications generated by 4 SOTA models, including GPT-3.5, and 2 humans.

Empirical experiments show that LENS achieves a higher correlation of 0.331 with human ratings on SIMPEVAL$_{2022}$, which is more than twice as high as the correlation scores of 0.112 and 0.149 by BERTScore and SARI, respectively. We further demonstrate that incorporating LENS into decoding process, using minimum Bayes risk framework (Fernandes et al., 2022), can directly improve the automatic text simplification system's performance. We expect that our data collection method, including RANK & RATE, a list-wise human evaluation interface, can be easily adapted to other text generation tasks.

2 Background

Issues of Existing Automatic Metrics. SARI (Xu et al., 2016) is the most commonly used metric for text simplification that computes F1/precision scores of the n-grams inserted, deleted, and kept when compared to human references. As SARI measures n-gram string overlap, it penalizes simplifications that are synonymous to the reference but uses different words (see first two examples in Figure 1). Alva-Manchego et al. (2021) showed that BERTScore (Zhang et al., 2020), which measures similarity based on BERT (Devlin et al., 2019) embeddings, is better at capturing semantic similarity between the simplification system's outputs and the references. However, it fails to penalize conservative systems that make trivial or no changes to the input as illustrated in Figure 2. This figure shows average metric scores by SARI and BERTScore for a conservative system that copies the input (*Copy*), an oracle system of human simplification (*Human*)[1], and three disfluent corrupted systems that drop 10% of the input words (*Drop*), scramble 5% of the input words (*Scramble*), or insert a period in the middle (*Split*). BERTScore ranks some conservative or corrupted systems above human simplifications. SARI penalizes the conservative system because it focuses on the edits performed by the system but ranks the disfluent systems above the conservative system. An ideal automatic metric for text simplification should capture both semantic similarity and the edits performed by the system.

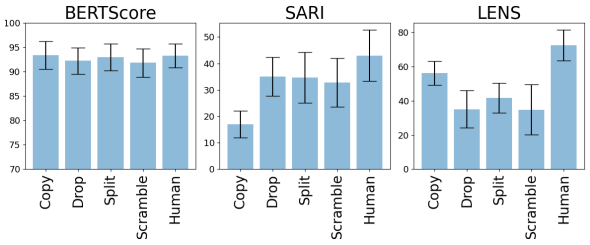


Figure 2: Average metric scores of conservative (*Copy*) or corrupted (*Drop*, *Scramble*, and *Split*) systems and human (*Human*) simplifications from SIMPEVAL$_{2022}$. The error bar represents the standard deviation of the metric scores. BERTScore ranks some conservative or corrupted systems above *Human*. SARI penalizes *Copy* but ranks the corrupted systems above *Copy*. LENS correctly ranks *Human* above the rest and *Copy* above the disfluent corrupted systems.

Lack of Human Evaluation Data. There is a lack of suitable human evaluation datasets for fine-tuning pre-trained language models for simplification evaluation. Alva-Manchego et al. (2021) released 0-100 continuous scale ratings on fluency, meaning preservation, and simplicity for only 600 simplifications sampled from six systems. Sulem et al. (2018) released 5-point Likert scale ratings on the same three dimensions for 1,960 simplifications generated by different configurations of six systems. Both datasets are relatively small and cover too few different system designs. For example, they do not include the newer state-of-the-art systems (Sheang and Saggion, 2021; Martin et al., 2022) that are based on T5 (Raffel et al., 2020) and other large pre-trained language models.

[1]We use SIMPEVAL$_{2022}$ (§3.1) for the analysis. Out of the two human simplifications, we randomly selected one as the oracle and the other as the reference to compute metric scores.

3 Automatic Evaluation Metric

To tackle the challenge of limited human evaluation data, we curate SIMPEVAL, a corpus containing over 13K human judgements on 2.8K simplification from 26 systems. This facilitates the training and evaluation of LENS (A **L**earnable **E**valuatio**n** Metric for Text **S**implification), the first supervised automatic metric for text simplification evaluation. In this section, we first describe the creation of SIMPEVAL datasets in §3.1 and then LENS in §3.2.

3.1 Collecting Human Judgements

We collect SIMPEVAL$_{\text{PAST}}$, containing 12K human ratings on 2.4K simplifications from 24 systems on sentences from TurkCorpus (Xu et al., 2016), to train LENS. For evaluating LENS and other simplification metrics, we create SIMPEVAL$_{2022}$ that consists of 1,080 human ratings on 360 simplifications from both humans and SOTA models, including GPT-3.5. It features more complex sentences from Wikipedia written after Oct 22nd, 2022, very recent to the time we conduct the experiments, to reduce the risk of "data contamination" (i.e., appearing in the training data of LLMs) and serve as a more challenging test bed for large language models. Table 1 shows the summary of both datasets.

A Diverse Set of Simplification Systems. We consider the following systems (further details in Appendix C): (i) two GPT-3.5[2] outputs under zero-shot and 5-shot settings; (ii) eight fine-tuned Transformer-based systems of varied sizes and parameter settings (Sheang and Saggion, 2021; Raffel et al., 2020; Martin et al., 2020; Maddela et al., 2021); (iii) three supervised BiLSTM-based systems that use vanilla RNN, reinforcement learning (Zhang and Lapata, 2017), or explicit editing (Dong et al., 2019); (iv) one unsupervised and one semi-supervised system utilizing auto-encoders (Surya et al., 2019); (v) two systems that apply statistical machine translation approaches to simplification (Wubben et al., 2012; Xu et al., 2016); (vi) a rule-based system (Kumar et al., 2020); (vii) three hybrid systems that combine linguistic rules with data-driven methods (Narayan and Gardent, 2014; Sulem, 2018; Maddela et al., 2021); (viii) two naive baselines that copy the input or scramble 5% of the input words; and (ix) six human-written simplifications, including two from ASSET (Alva-Manchego et al., 2020), one from TurkCorpus (Xu et al., 2016), one from Simple Wikipedia that were automatically aligned (Kauchak, 2013), and two newly written by our trained in-house annotators.

[2]Specifically, we use the text-davinci-003 model which is the most recent variant with 175B parameters. We provide the prompts we used in Appendix D.2.

System	Arch.	Data	Human Avg. Raw	Human Avg. Z-Score
SIMPEVAL$_{\text{PAST}}$ *(24 systems on 100 original sentences)*				
Human-1 (2020)	Human	ASSET	86.69	0.783
Human-2 (2020)	Human	ASSET	86.12	0.711
MUSS (2022)	BART-large	WikiLarge + Mined Data	84.48	0.653
ControlT5 (2021)	T5-base	WikiLarge	84.70	0.650
T5-3B (2020)	T5	WikiAuto	82.79	0.492
T5-large	T5	WikiAuto	81.86	0.453
T5-base	T5	WikiAuto	82.15	0.443
Transformer (2017)	BERT-TF	WikiAuto	79.42	0.366
Controllable (2021)	BERT-TF	WikiAuto	79.44	0.323
Human-3 (2016)	Human	TurkCorpus	79.54	0.281
Human-4	Human	Simple Wiki	78.36	0.249
DRESS (2017)	BiLSTM+RL	WikiLarge	77.18	0.206
Copy	–	–	76.81	0.103
ACCESS (2020)	Transformer	WikiLarge	73.25	0.001
SBMT-SARI (2016)	Statistic MT	PWKP	73.66	-0.014
PBMT-R (2012)	Statistic MT	PWKP	72.44	-0.066
EditNTS (2019)	BiLSTM	WikiLarge	70.15	-0.162
BiLSTM	BiLSTM	WikiLarge	68.35	-0.245
SEMosses (2018)	LSTM	WikiLarge	62.84	-0.565
UNTS (2019)	RNN	Wiki dump	62.66	-0.596
UNMT (2018)	RNN	Wiki dump	60.43	-0.673
Rule-based (2020)	–	–	60.15	-0.687
Hybrid (2014)	Statistic MT	PWKP	55.36	-0.925
Scramble	–	–	35.17	-1.954
SIMPEVAL$_{2022}$ *(6 systems on 60 original sentences)*				
Human-5	Human	Ours	81.87	0.424
Human-6	Human	Ours	81.57	0.395
GPT-3.5 (2022)	InstructGPT	5-shot	79.59	0.280
GPT-3.5 (2022)	InstructGPT	0-shot	75.27	-0.025
MUSS (2022)	BART-large	See above	70.74	-0.333
T5-3B (2020)	T5	WikiAuto	64.98	-0.700

Table 1: SIMPEVAL$_{\text{PAST}}$ and SIMPEVAL$_{2022}$ datasets of human evaluation data that covers a wide range of simplification systems. MUSS is the best-performing model in SIMPEVAL$_{\text{PAST}}$ with a small gap to humans but is much worse on the more challenging sentences in SIMPEVAL$_{2022}$ where GPT-3.5 performs better. BERT-TF: BERT-base initialized Transformer. Z-scores (Graham et al., 2013; Akhbardeh et al., 2021) are standardized based on each rater's mean and standard deviation.

Complex Sentences Selection. For SIMPEVAL$_{\text{PAST}}$, we sample 100 complex sentences from the test set of the widely-used TurkCorpus (Xu et al., 2016) and ASSET (Alva-Manchego et al., 2020) evaluation benchmarks, which share the same set of complex sentences but have different human simplifications in terms of edit operations. ASSET contains 10 human references for each complex sentence that are used to train LENS. GPT is pre-trained on a vast amount of web text, and

the sentences in TurkCorpus/ASSET are derived from the 12-year-old Parallel Wikipedia Simplification corpus (Zhu et al., 2010), which may have already been encountered by GPT-3.5. To address this "data contamination" issue and provide a more challenging benchmark for state-of-the-art models, we manually select 60 long, complex sentences covering recent world events such as the Twitter acquisition and World Cup for SIMPEVAL$_{2022}$.[3] These sentences are from new revisions or articles added to Wikipedia between October 22nd and November 24th, 2022. They have an average length of 38.5 tokens, much longer than the 19.7-token average of TurkCorpus and ASSET.

Data Annotation. We ask in-house annotators to rate each simplification on a single 0-100 overall quality scale using our RANK & RATE framework (more details in §6). We provided an extensive tutorial and two training sessions to the annotators, which involved rating system outputs for five input sentences (screenshots in Appendix F). We periodically inspected the annotations to prevent the deterioration of quality over time. All annotators are English speakers who are university students. Each simplification in SIMPEVAL$_{PAST}$ receives 5 ratings, while each simplification in SIMPEVAL$_{2022}$ is rated by 3 annotators. We follow WMT (Akhbardeh et al., 2021) to normalize the raw scores by the mean and standard deviation of each rater, also known as the z-score, which are later used to train LENS. The inter-annotator agreement for SIMPEVAL$_{PAST}$ is 0.70 using the interval Krippendorff's α on z-scores, and 0.32 for SIMPEVAL$_{2022}$, partly because it contains GPT-3.5 outputs that are quite competitive with human. Both are considered fair to good agreement (Krippendorff, 2004).

3.2 A New Learnable Metric – LENS

Given an input text c, the corresponding system output s, and a set of n references $R = \{r_1, r_2, \ldots, r_n\}$, LENS produces a real-valued score $z_{max} = \max_{1 \leq i \leq n}(z_i)$ that maximizes over the quality scores z_i of s in regards to each reference r_i. Our model encodes all texts into vectors $(\mathbf{c}, \mathbf{s}, \mathbf{r_i})$ using Transformer-based encoders such as RoBERTa (Liu et al., 2019), then combines them into an intermediate representation $\mathbf{H} = \left[\mathbf{s}; \mathbf{r_i}; \mathbf{s} \odot \mathbf{c}; \mathbf{s} \odot \mathbf{r_i}; |\mathbf{s} - \mathbf{c}|; |\mathbf{s} - \mathbf{r_i}|\right]$ by concatenation ($[;]$) and element-wise product ($\odot$), which is then fed to a feedforward network to predict z_i.

For training, besides considering all references equally (i.e., LENS$_{all}$ when $k = n$ in Eq. (1)), we also adopt a reference-adaptive loss that selects a subset of references closer to s in terms of edit operations rather than the entire set R. It encourages the metric to consider that different simplifications (e.g., paraphrasing-focused, deletion-focused, with or without splitting) can be acceptable, as long as they are close to some (not necessarily all) of the human references. We compute this loss (L_{adapt}) as:

$$L_{adapt} = \frac{1}{km} \sum_{j=1}^{m} \sum_{z_l \in Z'} (h_j - z_l)^2 \qquad (1)$$

where h is human rating and m is the training data size. We compute the set of predicted scores $Z = \{z_1, z_i, \ldots, z_n\}$ corresponding to references in R and then choose top k ($k \leq n$) values from Z to form a subset $Z' \subseteq Z$. Finally, we calculate the mean squared error (MSE) between the human rating h and each score in Z'. By selecting top k scores in Z, we focus the training of metric on references similar to s in terms of writing style or editing operations. This loss also aligns the training step with the inference step, where we select the best-predicted score z_{max} corresponding to R. Although multiple references are ideal, the proposed loss can also use a single reference, similar to the standard MSE loss. We train LENS on our SIMPEVAL$_{PAST}$ dataset (details in §3.1). We provide further implementation details in Appendix A.

Similar to the COMET (Rei et al., 2020) and BLEURT (Sellam et al., 2020) metrics for MT evaluation, LENS is trained on human ratings after z-score normalization, which are real values predominantly lie between [-3, 3]. To make LENS more interpretable, we rescale the predicted scores to the range between [0, 100] as the percentage of area under a normal curve of mean 0 and standard deviation 1 corresponding to z_i. We present the rescaled scores for experiments and analyses in this paper. Besides, we use RoBERTa-large as the underlying model of LENS throughout the paper, unless otherwise specified (§5).

4 Experiments

We benchmark LENS metric against the existing automatic metrics (i) for evaluating text simplification system outputs and (ii) for training better

[3]For example, *"Musk stated that Twitter Blue's pricing would be raised to around US$8.00 per-month, and include reduced advertising on the Twitter service, the ability to post longer audio and video files, and verified account status."*

	SIMPEVAL$_{2022}$			WIKI-DA			NEWSELA-LIKERT		
	τ_{para}	τ_{spl}	τ_{all}	**Fluency**	**Meaning**	**Simplicity**	**Fluency**	**Meaning**	**Simplicity**
FKGL	-0.556	-0.31	-0.356	0.054	0.145	0.001	0.193	0.306	-0.051
BLEU	0.048	-0.054	-0.033	0.460	0.622	0.438	0.332	0.261	0.118
SARI	0.206	0.140	0.149	0.335	0.534	0.366	0.234	0.124	0.094
BERTScore	0.238	0.093	0.112	0.636	**0.682**	0.614	0.384	0.274	0.215
LENS$_{all}$	0.333	0.233	0.241	**0.816**	0.662	0.733	**0.655**	**0.477**	0.343
LENS$_{k=1}$	**0.460**	0.295	0.307	0.796	0.647	0.721	0.633	0.444	0.328
LENS$_{k=3}$	0.429	**0.333**	**0.331**	0.807	0.668	**0.749**	0.624	0.428	**0.359**

Table 2: Correlation results between automatic metrics and three human ratings datasets: SIMPEVAL$_{2022}$ (this work), WIKI-DA (Alva-Manchego et al., 2021), and NEWSELA-LIKERT (Maddela et al., 2021). τ_{para}, τ_{spl}, and τ_{all} represent the Kendall Tau-like correlation for paraphrase-focused, split-focused, and all simplifications, respectively. We report the Pearson correlation coefficients along three dimensions for WIKI-DA and NEWSELA-LIKERT. The best values are marked in **bold** and the second best values are underlined.

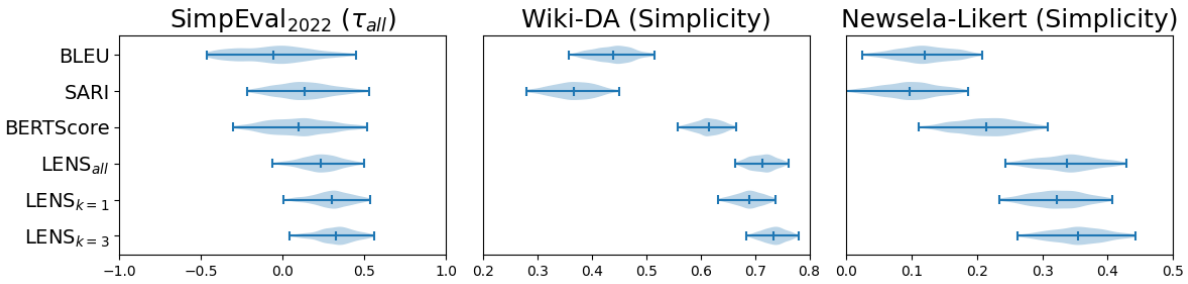


Figure 3: The 95% confidence intervals for Kendall Tau-like correlation (τ_{all}) on SIMPEVAL$_{2022}$ and for Pearson correlation with simplicity ratings on WIKI-DA and NEWSELA-LIKERT, calculated by bootstrapping (Deutsch et al., 2021). LENS is more reliable with smaller intervals and has higher correlation with human judgments.

automatic simplification models when used as alternative reward functions.

4.1 Correlation with Human Evaluation

We demonstrate that LENS correlates better with human judgments than the existing metrics.

Evaluation Datasets. We compare LENS to the existing metrics, namely SARI (Xu et al., 2016), BERTScore (Zhang et al., 2020),[4] BLEU, and Flesch-Kincaid Grade Level readability (FKGL) (Kincaid, 1975) on three datasets: SIMPEVAL$_{2022}$ (§3.1), WIKI-DA released by Alva-Manchego et al. (2021) with 0-100 continuous scale ratings on fluency, meaning preservation, and simplicity for 600 simplifications across six systems, and NEWSELA-LIKERT collected by Maddela et al. (2021) with 5-point Likert scale ratings on the same dimensions for 500 simplifications across five systems. While SIMPEVAL$_{2022}$ and WIKI-DA are derived from Wikipedia, NEWSELA-LIKERT is derived from news articles in Newsela (Xu et al., 2015), a widely used corpus for text simplification. When calculating metric scores for SIMPEVAL$_{2022}$ that contains two human simplifications, we use one as the reference and the other as the oracle simplification system. We remove the complex sentences in WIKI-DA that overlap with SIMPEVAL$_{PAST}$, the training dataset for LENS.

Evaluation Setup. We report Kendall Tau-like correlation for SIMPEVAL$_{2022}$ to capture the ability of metrics to distinguish two systems, which is close to the real-world scenario. Kendall Tau-like correlation is predominantly used in machine translation metric evaluation at WMT (Bojar et al., 2017; Ma et al., 2018) for the same reason as it focuses on the relative ranking of the outputs. Given an input c and its simplifications from N systems $S = \{s_1, \ldots, s_m, \ldots, s_n, \ldots, s_N\}$, we extract (s_m, s_n) pairs, where $1 \leq m < n \leq N$, and calculate Kendall Tau-like coefficient τ:

$$\tau = \frac{|Concordant| - |Discordant|}{|Concordant| + |Discordant|} \quad (2)$$

where *Concordant* is the set of pairs where the metric ranked (s_m, s_n) in the same order as humans ranked and *Discordant* is the set of the pairs where the metric and humans disagreed. We report separate coefficients for paraphrase- ($\tau_{\textbf{para}}$),

[4]We use precision score of BERTScore as it has been shown to perform better for simplification evaluation (Alva-Manchego et al., 2021).

and split-focused ($\tau_{\textbf{spl}}$) simplifications, along with ($\tau_{\textbf{all}}$) for all of them. Following the literature, we only used the (s_m, s_n) pairs for which the difference in ratings is more than 5 out of the 100 points, and all the annotators agreed on the ranking order. To make results comparable to existing work (Alva-Manchego et al., 2021, 2020) that evaluated on WIKI-DA and NEWSELA-LIKERT, we report Pearson correlation (ρ) between the metric scores and the human ratings.

Results. Table 2 shows that LENS outperforms the existing metrics on SIMPEVAL$_{2022}$ and WIKI-DA belonging to the Wikipedia domain, and NEWSELA-LIKERT based on newswire domain. The difference is more substantial on SIMPEVAL$_{2022}$ consisting of similar performing SOTA systems, where the τ_{all} of LENS$_{k=3}$ exceeds the τ_{all} of BERTScore by 0.22 points. Training using top k references (LENS$_{k=3}$) has improved τ_{all} on SIMPEVAL$_{2022}$ and ρ along the simplicity dimension on the rest when compared to using all the references (LENS$_{all}$). Figure 3 shows the 95% confidence intervals for τ_{all} on SIMPEVAL$_{2022}$ and ρ for simplicity, which is deemed to be the most important dimension in prior work (Alva-Manchego et al., 2021; Xu et al., 2016), on WIKI-DA and NEWSELA-LIKERT calculated using bootstrapping methods by Deutsch et al. (2021). A smaller interval indicates that the metric exhibits lower variance and higher reliability. LENS exhibits smaller intervals than the other metrics.

4.2 LENS as Training or Decoding Objectives

We also incorporate LENS into training as an alternative reward function in minimum risk training framework (Kumar and Byrne, 2004; Smith and Eisner, 2006) and into decoding as a reranking objective in minimum Bayes risk decoding (Fernandes et al., 2022; Freitag et al., 2022) to improve the quality of generated simplifications.

Minimum Risk Training (MRT). Given the input c and reference r, we generate a set of candidates S for each training step and calculate the expected risk (L_{risk}) as follows:

$$L_{risk} = \sum_{s \in S} cost(c, r, s) \frac{P(s|c)}{\sum_{s' \in S} P(s'|c)} \quad (3)$$

$$P(s|c) = \prod_{t=1}^{T} P(s_t|c, s_{<t}; \theta)$$

$$cost(c, r, s) = 1 - Metric(c, r, s)$$

	BL	SARI	BS	LENS	sBL ↓	Human
MLE	44.4	47.5	**94.7**	61.0	70.7	68.71
Minimum Risk Training (MRT)						
BL	**45.1**	47.3	**94.7**	60.5	71.3	68.33
SARI	43.4	48.5	94.5	63.2	68.1	70.86
BS	44.0	48.4	94.6	62.1	67.8	70.32
LENS	42.6	47.4	94.5	63.9	67.1	68.89
Minimum Bayes Risk Decoding (MBR)						
BL	44.8	48.3	94.4	60.4	74.1	66.19
SARI	43.9	**49.6**	94.5	59.3	75.5	64.31
BS	44.3	48.3	94.6	61.1	73.1	66.79
LENS	43.2	49.5	**94.7**	**73.0**	**61.8**	**72.80**

Table 3: Evaluation results for minimum risk training (MRT) and minimum Bayes risk decoding (MBR) using T5-base model on SIMPEVAL$_{2022}$. We report **LENS**, BLEU (**BL**), **SARI**, BERTScore (**BS**), self-BLEU (**sBL**), and average human ratings. MBR with LENS shows the best human evaluation results despite making the most number of edits to the input as indicated by its lowest self-BLEU. We use beam search with beam size of 10 for MLE.

where $Metric(c, r, s)$ can be any evaluation metric, θ are the model parameters, and $s_{<t} = s_1, \ldots s_{t-1}$ is a partial generation of s.

Following previous work (Shen et al., 2016; Wieting et al., 2019), we first train the seq-to-seq generation model with maximum likelihood estimation (MLE) for a few epochs and then we change the loss to a combination of L_{MLE} and expected risk:

$$L_{MRT} = \gamma L_{MLE} + (1 - \gamma) L_{risk}. \quad (4)$$

We choose $\gamma = 0.3$ and $|S| = 10$ for our experiments.

Minimum Bayes Risk Decoding (MBR). We adopt the MBR framework proposed by Fernandes et al. (2022), where we first generate a set of candidates S for input c during inference then rerank them by comparing each candidate $s \in S$ to all the other candidates in the set:

$$\hat{u}_{MBR} = \arg\max_{s \in S} \frac{1}{|S|} \sum_{s' \in S} Metric(c, s', s). \quad (5)$$

For our experiments, we generate the candidates using beam search with beam size $= |S|$.

Experiment Setup. We fine-tune a T5 model that prepends control tokens to the input (Sheang and Saggion, 2021) to control various aspects of the generated simplification and has shown state-of-the-art performance for the task. We use WIKI-AUTO (Jiang et al., 2020) for training, ASSET

	BL	SARI	BS	LENS	sBL ↓
T5-base					
$\text{MLE}_{b=10}$	**44.4**	47.5	94.7	61.0	70.7
$\text{MLE}_{b=100}$	42.6	46.2	94.3	57.6 (-3.4)	68.1
$\text{MBR-LENS}_{\lvert S\rvert=10}$	43.6	48.7	94.6	67.8 (+6.8)	67.0
$\text{MBR-LENS}_{\lvert S\rvert=100}$	43.2	49.5	94.7	73.0 (+12.1)	61.8
T5-3B					
$\text{MLE}_{b=10}$	42.3	46.3	94.8	60.3	61.9
$\text{MLE}_{b=100}$	39.7	42.5	94.0	53.4 (-6.9)	65.4
$\text{MBR-LENS}_{\lvert S\rvert=10}$	41.3	46.1	94.7	66.9 (+6.6)	58.8
$\text{MBR-LENS}_{\lvert S\rvert=100}$	42.3	**47.7**	**94.9**	72.6 (+12.3)	55.7
T5-11B					
$\text{MLE}_{b=10}$	37.6	46.4	93.8	62.9	49.3
$\text{MLE}_{b=100}$	35.8	44.7	93.2	59.2 (-3.7)	51.1
$\text{MBR-LENS}_{\lvert S\rvert=10}$	36.9	46.5	93.9	69.9 (+7.0)	48.4
$\text{MBR-LENS}_{\lvert S\rvert=100}$	36.2	46.1	93.8	**74.4** (+11.5)	**44.6**

Table 4: Automatic evaluation results for minimum Bayes risk decoding (MBR) with different model sizes on SIMPEVAL_{2022}. For standard MLE, we use beam search with beam sizes of 10 and 100.

	BL	SARI	BS	LENS	sBL↓	Human
T5-11B						
$\text{MLE}_{b=10}$	**37.6**	**46.4**	93.8	62.9	49.3	88.80
$\text{MBR-LENS}_{\lvert S\rvert=100}$	36.2	46.1	93.8	**74.4**	44.6	90.13
Close-source LLMs						
GPT-3.5 (0-shot)	27.8	41.4	93.4	60.7	31.8	90.77
GPT-3.5 (5-shot)	30.5	42.4	94.1	69.0	33.2	92.70
GPT-4 (0-shot)	31.6	43.7	**94.3**	73.5	**29.1**	**93.63**

Table 5: Human evaluation results for the T5-11B and close-sourced LLMs on SIMPEVAL_{2022}. T5-11B with MBR-LENS decoding achieves the state-of-the-art open-source model performance, on par with GPT-3.5.

for validation, and the challenging, complex sentences in SIMPEVAL_{2022} for testing. WIKIAUTO consists of 400k complex-simple sentence pairs extracted from Normal and Simple Wikipedia document pairs. For our experiments, we trained T5 towards LENS, BLEU, SARI, and BERTScore. We provide more implementation details in Appendix B. We report the same metrics on the test set along with self-BLEU to capture the diversity of the outputs. We also ask 3 annotators to rate the system outputs from SIMPEVAL_{2022} using our evaluation interface (§6) and report the averaged ratings.

Results. Table 3 shows that LENS integrated into minimal Bayes risk decoding (MBR-LENS) achieves the best human evaluation results while it makes the most number of edits to the input (less copying) as indicated by its lowest self-BLEU score. Although MRT-LENS has slightly lower average human ratings than MRT-BERTScore and MRT-SARI, the pairwise comparison shows that

	#Param	τ_{para}	τ_{spl}	τ_{all}
MiniLM	66M	0.143	0.310	0.277
BERT-base	110M	0.461	0.240	0.258
BERT-large	340M	0.461	0.279	0.298
RoBERTa-base	110M	**0.472**	0.271	0.295
RoBERTa-large	340M	0.429	**0.333**	**0.331**

Table 6: Kendall Tau-like correlation (τ_{para} , τ_{spl}, and τ_{all}) of LENS metric, when based on different encoder models, with human ratings in SIMPEVAL_{2022} .

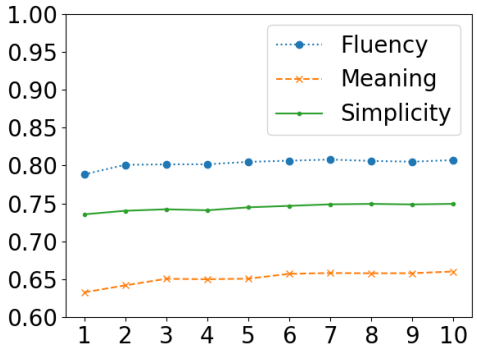


Figure 4: Pearson correlation of LENS on WIKIDA dataset using a varied number of references.

they are very comparable: MRT-LENS generates 28% better, 28% worse, and 44% equal quality simplifications compared to MRT-SARI. Additionally, BERTScore and SARI show a decrease in quality when used in decoding than standard maximum likelihood estimation (MLE). It is noteworthy that we use a beam size of 10 for MLE rather than a large search space of beam size 100 because generation quality degrades with increased beam size as shown in Table 4 as well as in existing literature (Stahlberg and Byrne, 2019; Meister et al., 2020).

Given the success of using LENS as utility function of MBR decoding on T5-base, we further apply it to larger models, including T5-3B and T5-11B. As displayed in Table 4, MBR-LENS with 100 candidates improves over standard beam search by an increase of over 11 points of LENS score across all model sizes. Although MBR may inflate the results of the utility function it uses (Fernandes et al., 2022), our human evaluations solidify the assertion that T5-11B with MBR-LENS decoding exceeds standard beam search, thereby establishing state-of-the-art (SOTA) performance among open-source models. When compared to close-source large language models, T5-11B with MBR-LENS achieves on-par performance with GPT-3.5.

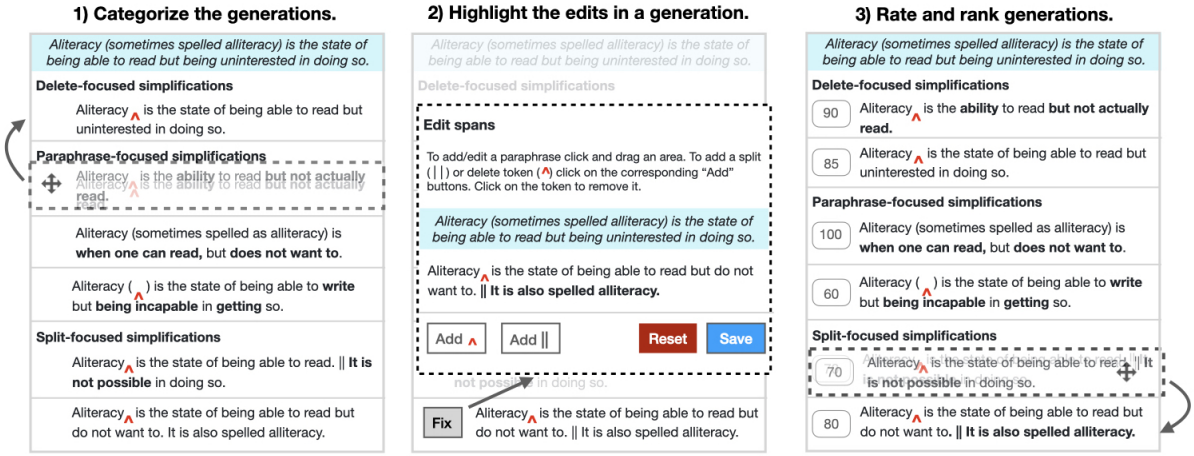


Figure 5: RANK & RATE framework consists of three steps: (1) classifying the generations, (2) annotating the edits performed by the system, and (3) rating and ranking the generations. For the first two steps, the annotators verify the automatically extracted categories and edits instead of annotating from scratch.

5 LENS Analysis

In this section, we delve into the impact of the underlying model architecture and the number of references on the performance of LENS.

Model Architecture. Table 6 shows the Kendall Tau-like correlation (τ_{para}, τ_{spl}, and τ_{all}) of LENS metric trained on various encoders on SIMPEVAL$_{2022}$. LENS metrics trained on RoBERTa encoders (Liu et al., 2019) perform better than their respective LENS metrics trained on the BERT encoders (Devlin et al., 2019). Among all models, LENS trained on RoBERTa-large achieves the highest overall correlation. It has a substantial improvement in τ_{spl} with a trade-off in τ_{para}, in comparison to RoBERTa-base.

Number of References. Figure 4 shows the Pearson correlation of LENS metric on WIKI-DA for a varied number of references used during inference. Although the metric performs the best with 10 references, we see only a slight drop with one reference, demonstrating that LENS is capable of evaluating with single or multiple references.

6 RANK & RATE Framework

To facilitate consistent evaluation and enable the training of LENS, we develop RANK & RATE , a human evaluation framework to assist annotators in efficiently comparing and rating many (>20) system outputs at once in a list-wise ranking manner.

6.1 Methodology

We describe the three-step annotation methodology for RANK & RATE (Figure 5) as follows:

Step 1 - Categorizing System Outputs. As there are different acceptable ways to simplify the input text, we first display the system outputs in groups as split-, deletion-, or paraphrase-focused based on the following criteria: (i) outputs with multiple sentences are split-focused, (ii) outputs with a compression ratio less than 0.5 or generated by deleting words from the original sentence are deletion-focused, (iii) the rest are paraphrase-focused. Annotators can adjust the initial automatic categorization during the annotation process.

Step 2 - Highlighting Edits. To help annotators notice the changes made by each system and subsequently improve the accuracy of their ratings, we use a state-of-the-art word alignment model by Lan et al. (2021) to highlight the edits performed by the system and classify them into three types: (i) Deletion-edits are phrases or words with no match in the modified output, (ii) Paraphrase-edits are new phrases or words added or modified, and (iii) Split-edits are any added periods (“.”). Deletion-edits are marked with a red caret (“^”), paraphrase-edits with bolded text, and split-edits with two vertical bars (“ || ”) (see Figure 5). We ask annotators to correct the misclassified edits using an interactive pop-up window (see Appendix F).

Step 3 - Ranking and Rating System Outputs. Following the machine translation evaluation at WMT (Ma et al., 2018, 2019; Barrault et al., 2020; Akhbardeh et al., 2021), we ask annotators to rate the quality of system outputs on a 0-100 continuous scale instead of the 5-point Likert scale because the former was shown to provide higher levels of inter-annotator consistency (Novikova et al., 2018) and

	SIMPLIKERT$_{2022}$				SIMPDA$_{2022}$				SIMPEVAL$_{2022}$
	Fluency	**Adequacy**	**Simplicity**	**Avg.**	**Fluency**	**Adequacy**	**Simplicity**	**Avg.**	**Overall**
Human-1	4.69	4.46	1.36	3.50	92.14	**88.87**	83.91	88.30	81.57
Human-2	**4.72**	4.48	**1.39**	**3.53**	92.84	88.23	**84.50**	88.53	**81.87**
GPT-3.5 (5-shot)	4.64	**4.66**	1.13	3.48	92.59	92.10	81.17	**88.62**	79.59
GPT-3.5 (zero-shot)	4.63	4.56	0.91	3.37	90.51	90.32	74.81	85.21	75.27
MUSS	4.40	4.11	0.86	3.12	87.99	80.57	73.12	80.56	70.74
T5-3B	4.70	4.33	0.55	3.19	**93.96**	85.22	58.44	79.21	64.98

Table 7: Model and human simplification quality under different human evaluation methods. Following Sulem et al. (2018), SIMPLIKERT$_{2022}$ uses a 1 to 5 scale for fluency and adequacy, and -2 to 2 for simplicity. SIMPDA$_{2022}$ rates on a continuous 0-100 scale. All three methods show similar rankings of systems. **Bold**: the best.

is more suitable to apply on many statistical models (Graham et al., 2013). Our interactive interface allows the annotators to move the system outputs up and down to rank them and compare similar quality outputs more easily by placing them together.

We provide the annotation instructions and screenshots of the interface in Appendix F.

6.2 Human Evaluation Comparison

We compare RANK & RATE with the existing human evaluation methods: 5-point Likert (Sulem et al., 2018) and Direct Assessment with a continuous scale of 0 to 100 (Alva-Manchego et al., 2021), which were both conducted on three dimensions: fluency, adequacy, and simplicity. For a fair comparison, we annotate the same set of simplifications using each method, resulting in SIMPLIKERT$_{2022}$ and SIMPDA$_{2022}$. Table 7 shows similar rankings of the systems by the three methods with very slight differences. We also calculate the interval Krippendorff's α (Krippendorff, 2011) for inter-annotator agreement. RANK & RATE achieves an α of 0.32, which is higher than the 0.23 α by Likert and 0.25 α by DA. All values are considered fair (Krippendorff, 2004).

7 Other Related Work

Text-to-text Generation Metrics. There is currently no single automatic metric to evaluate all the text-to-text generation tasks that revise text. SARI (Xu et al., 2016) is the main metric for text simplification and has been used by other generation tasks such as sentence splitting and fusion (Rothe et al., 2020; Kim et al., 2021), decontextualization (Choi et al., 2021), and scientific rewriting (Du et al., 2022). Style transfer tasks (Hu et al., 2017; Rao and Tetreault, 2018; Prabhumoye et al., 2018; Krishna et al., 2020; Xu et al., 2012; Ma et al., 2020) use different automatic metrics to measure each aspect of text: (i) text similarity metrics such as BLEU, METEOR (Lavie and Agarwal, 2007), and BERTScore to measure content preservation, (ii) text classification models (Sennrich et al., 2016; Luo et al., 2019; Krishna et al., 2022b) or embedding-based edit distance metrics such as MoverScore (Zhao et al., 2019; Mir et al., 2019) to evaluate target style, and (iii) perplexity or pseudo log likelihood (Salazar et al., 2020) to measure fluency.

Incorporating Evaluation Metrics into Training and Inference. Prior studies have improved MLE-trained generation systems with alternative training approaches that integrate evaluation metrics based on reinforcement learning (Ranzato et al., 2015; Li et al., 2016; Gong et al., 2019), minimum risk (Smith and Eisner, 2006; Shen et al., 2016), and ranking (Hopkins and May, 2011; Xu et al., 2016; Krishna et al., 2022a). Incorporating metrics into decoding has also been explored by reranking the generations using discriminative rankers (Shen et al., 2004; Lee et al., 2021), energy-based rankers (Bhattacharyya et al., 2021), and minimum risk (Kumar and Byrne, 2004; Freitag et al., 2022). We use minimum risk as it has been shown to help machine translation systems (Wieting et al., 2019; Fernandes et al., 2022; Amrhein and Sennrich, 2022).

8 Conclusion

We introduce LENS, the first supervised automatic metric for text simplification. We show that LENS exhibits higher human correlation than other automatic metrics. We also introduce RANK & RATE framework, which allows annotators to evaluate multiple systems' outputs at once in a list-wise manner. Using it, we create SIMPEVAL$_{PAST}$ to train LENS, and SIMPEVAL$_{2022}$ as a new metric evaluation benchmark for text simplification. We hope our metric, data, and framework will facilitate future research in text simplification evaluation.

Limitations

In this paper, we show that LENS shows better human correlation than other metrics on Wikipedia and news domains. Future research can further experiment and extend LENS to other domains, such as medical and children's books, as the preference for different simplification operations can vary depending on the domain and user. Additionally, our work focuses on sentence-level simplification, and future work can extend LENS to evaluating paragraph- and document-level simplification. SIMPEVAL dataset and LENS are also limited to the English language.

Ethics Statement

We used in-house annotators to collect human ratings in SIMPEVAL datasets and write simplifications in SIMPEVAL_{2022}. The annotators are university-level undergraduate and graduate students, including both native and non-native speakers of English. We did not collect any personal information from the annotators. We paid each annotator $15 per hour, which is above the US federal minimum wage. We ensured that the content shown to the annotators was not upsetting and let them know that they could skip the task if they felt uncomfortable at any point. We also let the annotators know the purpose of the collected data.

The original complex sentences in the SIMPEVAL datasets are from the publicly available Wikipedia. The simplifications are either from the existing work or human simplifications collected from our annotators. We used the author-released simplification outputs if they are available. For T5 (base, large, and 3B) systems, we trained our own simplification models using open-sourced code from the Hugging Face Transformers[5] library.

Acknowledgments

We thank Nghia T. Le, Tarek Naous, Yang Chen, and Chao Jiang as well as three anonymous reviewers for their helpful feedback on this work. We also thank Marcus Ma, Rachel Choi, Vishnesh J. Ramanathan, Elizabeth Liu, Alex Soong, Govind Ramesh, Ayush Panda, Anton Lavrouk, Vinayak Athavale, and Kelly Smith for their help with human evaluation. This research is supported in part by the NSF awards IIS-2144493 and IIS-2112633, ODNI and IARPA via the BETTER program (contract 2019-19051600004) and the HIATUS program (contract 2022-22072200004). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of NSF, ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

[5]https://huggingface.co/

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

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 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.

Chantal Amrhein and Rico Sennrich. 2022. Identifying weaknesses in machine translation metrics through minimum Bayes risk decoding: A case study for COMET. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. *ArXiv*, abs/1710.11041.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi,

Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Sumanta Bhattacharyya, Amirmohammad Rooshenas, Subhajit Naskar, Simeng Sun, Mohit Iyyer, and Andrew McCallum. 2021. Energy-based reranking: Improving neural machine translation using energy-based models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4528–4537, Online. 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.

Eunsol Choi, Jennimaria Palomaki, Matthew Lamm, Tom Kwiatkowski, Dipanjan Das, and Michael Collins. 2021. Decontextualization: Making sentences stand-alone. *Transactions of the Association for Computational Linguistics (TACL)*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2021. A statistical analysis of summarization evaluation metrics using resampling methods. *Transactions of the Association for Computational Linguistics*, 9:1132–1146.

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

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

Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith, and Yejin Choi. 2022. Is GPT-3 text indistinguishable from human text? scarecrow: A framework for scrutinizing machine text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7250–7274, Dublin, Ireland. Association for Computational Linguistics.

Wanyu Du, Vipul Raheja, Dhruv Kumar, Zae Myung Kim, Melissa Lopez, and Dongyeop Kang. 2022. Understanding iterative revision from human-written text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3573–3590, Dublin, Ireland. Association for Computational Linguistics.

Patrick Fernandes, António Farinhas, Ricardo Rei, José De Souza, Perez Ogayo, Graham Neubig, and Andre Martins. 2022. Quality-aware decoding for neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412, Seattle, United States. Association for Computational Linguistics.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2022. High quality rather than high model probability: Minimum Bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825.

Hongyu Gong, Suma Bhat, Lingfei Wu, JinJun Xiong, and Wen-mei Hwu. 2019. Reinforcement learning based text style transfer without parallel training corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3168–3180, Minneapolis, Minnesota. Association for Computational Linguistics.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of GPT-3. *arXiv preprint arXiv:2209.12356*.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Mark Hopkins and Jonathan May. 2011. Tuning as ranking. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1352–1362, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning*.

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 2020. Neural CRF model for

sentence alignment in text simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7943–7960, Online. Association for Computational Linguistics.

David Kauchak. 2013. Improving text simplification language modeling using unsimplified text data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1537–1546, Sofia, Bulgaria. Association for Computational Linguistics.

Joongwon Kim, Mounica Maddela, Reno Kriz, Wei Xu, and Chris Callison-Burch. 2021. BiSECT: Learning to split and rephrase sentences with bitexts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6193–6209, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kincaid. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. *research branch report*.

Klaus Krippendorff. 2004. Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research*, 30(3):411–433.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Kalpesh Krishna, Yapei Chang, John Wieting, and Mohit Iyyer. 2022a. Rankgen: Improving text generation with large ranking models.

Kalpesh Krishna, Deepak Nathani, Xavier Garcia, Bidisha Samanta, and Partha Talukdar. 2022b. Few-shot controllable style transfer for low-resource multilingual settings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7439–7468, Dublin, Ireland. Association for Computational Linguistics.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 737–762, Online. Association for Computational Linguistics.

Dhruv Kumar, Lili Mou, Lukasz Golab, and Vechtomova Olga. 2020. Iterative edit-based unsupervised sentence simplification. In *Proceedings of the Association for Computational Linguistics*.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176, Boston, Massachusetts, USA. Association for Computational Linguistics.

Wuwei Lan, Chao Jiang, and Wei Xu. 2021. Neural semi-markov CRF for monolingual word alignment.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

Ann Lee, Michael Auli, and Marc'Aurelio Ranzato. 2021. Discriminative reranking for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7250–7264, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Austin, Texas. Association for Computational Linguistics.

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. *ArXiv*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Fuli Luo, Peng Li, Pengcheng Yang, Jie Zhou, Yutong Tan, Baobao Chang, Zhifang Sui, and Xu Sun. 2019. Towards fine-grained text sentiment transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2013–2022, Florence, Italy. Association for Computational Linguistics.

Qingsong Ma, Ondřej Bojar, and Yvette Graham. 2018. Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 671–688, Belgium, Brussels. Association for Computational Linguistics.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics

shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

Xinyao Ma, Maarten Sap, Hannah Rashkin, and Yejin Choi. 2020. PowerTransformer: Unsupervised controllable revision for biased language correction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7426–7441, Online. 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.

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

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

Clara Meister, Ryan Cotterell, and Tim Vieira. 2020. If beam search is the answer, what was the question? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2173–2185, Online. Association for Computational Linguistics.

Remi Mir, Bjarke Felbo, Nick Obradovich, and Iyad Rahwan. 2019. Evaluating style transfer for text. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 495–504, Minneapolis, Minnesota. 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.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2018. RankME: Reliable human ratings for natural language generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 72–78, New Orleans, Louisiana. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Richard Yuanzhe Pang and Kevin Gimpel. 2019. Unsupervised evaluation metrics and learning criteria for non-parallel textual transfer. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 138–147, Hong Kong. Association for Computational Linguistics.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–876, Melbourne, Australia. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. 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 129–140, New Orleans, Louisiana. 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.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280.

Horacio Saggion. 2017. Automatic text simplification. *Synthesis Lectures on Human Language Technologies*.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. 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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, San Diego, California. Association for Computational Linguistics.

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

Libin Shen, Anoop Sarkar, and Franz Josef Och. 2004. Discriminative reranking for machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 177–184, Boston, Massachusetts, USA. Association for Computational Linguistics.

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Minimum risk training for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1692, Berlin, Germany. Association for Computational Linguistics.

David A. Smith and Jason Eisner. 2006. Minimum risk annealing for training log-linear models. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 787–794, Sydney, Australia. Association for Computational Linguistics.

Felix Stahlberg and Bill Byrne. 2019. On NMT search errors and model errors: Cat got your tongue? 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 3356–3362, Hong Kong, China. Association for Computational Linguistics.

Elior Sulem. 2018. 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)*.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. 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.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*.

John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019. Beyond BLEU:training neural machine translation with semantic similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4344–4355, Florence, Italy. Association for Computational Linguistics.

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

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

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.

Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, and Colin Cherry. 2012. Paraphrasing for style. In *Proceedings of COLING 2012*, pages 2899–2914, Mumbai, India. The COLING 2012 Organizing Committee.

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.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. 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 563–578, Hong Kong, China. Association for Computational Linguistics.

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 LENS Metric - Implementation Details

We leverage *Transformers* and *PyTorch Lightning* libraries to implement the metric. We used RoBERTa$_{Large}$ model (Liu et al., 2019) as the encoder. We fine-tuned the metric with the style-adaptive loss for 20 epochs and selected the checkpoint with the best validation loss. We used the ten human references from ASSET to calculate metric scores while training LENS using SIMPEVAL$_{PAST}$. We used a batch size of 8, Adam optimizer with a learning rate of 3e-05 for the encoder and 1e-05 for other layers. We use dropout of 0.15 and freeze the encoding layers for 1 epoch. We train the model on one Quadro RTX 6000 GPU with 25GB RAM, which takes around 3 hours.

B Incorporating Evaluation Metrics into Training and Inference - Implementation Details

We implement the controllable sentence simplification system proposed by Sheang and Saggion (2021), a T5-base model. The input sentence is prepended with four control tokens, namely character length ratio, dependency tree depth ratio, character-level Levenshtein similarity, and inverse frequency ratio, to control various aspects of the generated simplification. During training, each control value is calculated using the corresponding training pair and discretized into bins of width 0.05. During inference, we set the control tokens to the average value of the training set. We used 0.9 for the length ratio and 0.75 for the rest. We also fine-tune the controllable T5-3B and T5-11B versions using the same approach.

We use the Hugging Face Transformers library[6] for implementing the base model and the MRT framework. We first train the model for 5 epochs using MLE loss and then fine-tune it for 5 epochs using MRT. We use Adam optimizer with a learning rate of 1e-4, linear learning rate warmup of 4k steps, weight decay of 0.1, epsilon of 1e-8, and batch size of 16. During MRT, we use a beam size of 8 to generate candidates. The rest of the parameters are left with default values from the library. During inference, we use a beam size of 10. Our models are trained on 2 A40 GPUs with 45GB RAM for 48 hours.

We used the code released by Fernandes et al. (2022)[7] for the MBR framework. We generated candidates using beam search of size 100 and selected the top 100 candidates for reranking. We used the above T5 model fine-tuned using MLE to generate the candidates. The rest of the parameters are left with default values from the library.

C Systems in SIMPEVAL Datasets

Data-driven Neural Models. We use ten supervised models: (i) three fine-tuned T5 (Raffel et al., 2020) models of various sizes, namely T5-base, T5-large, and T5-3B, (ii) a controllable T5-base model that prepends tokens to the input to control the lexical and syntactic complexity of the output (Sheang and Saggion, 2021), (iii) two BART (Lewis et al., 2020) models that also use control tokens where one is trained on Wikipedia (Martin et al., 2020), and the other is fine-tuned on a combination of Wikipedia and web-mined paraphrases (Martin et al., 2022), (iv) a BERT-initialized Transformer (Maddela et al., 2021), (v) a BiLSTM edit-based approach that first generates the edit operations, and then the simplification (Dong et al., 2019), (vi) a BiLSTM that directly generates simplifications using reinforcement learning (Zhang and Lapata, 2017), and (vii) a vanilla BiLSTM model. In addition, we also include one unsupervised model and one semi-supervised model by Surya et al. (2019) that uses an auto-encoder with adversarial and denoising losses.

Few-shot Methods. We include simplifications generated by GPT-3.5[8] under zero-shot and 5-shot settings (prompts are provided in Appendix D.2).

Data-driven Statistical Methods. We incorporate two systems that applied statistical machine translation (MT) approaches to text simplification: (i) a phrase-based MT system that reranks the outputs based on their dissimilarity with the input (Wubben et al., 2012), and (ii) a syntactic MT system that uses paraphrase rules for lexical simplification (Xu et al., 2016).

Rule-based Methods. Kumar et al. (2020) iteratively generates candidates using rules and ranks them with a linguistic scoring function.

Hybrid Methods. We utilize three hybrid systems that combine linguistic rules with data-driven methods: (i) Narayan and Gardent (2014) uses semantic structure to predict sentence splitting and

[6] https://huggingface.co/

[7] https://github.com/deep-spin/qaware-decode

[8] Specifically, we use the text-davinci-003 model which is the most recent variant with 175B parameters.

paraphrases with a phrase-based MT system, (ii) Sulem et al. (2018) performs splitting and deletion using linguistic rules and paraphrasing using a BiLSTM, and (iii) Maddela et al. (2021) generates candidates with different amounts of splitting and deletion and then paraphrases the best candidate with a BERT-initialized Transformer.

Naive Baseline Methods. Existing metrics are biased towards conservative systems because their outputs are generally fluent and exhibit high lexical overlap with the input/reference (Pang and Gimpel, 2019; Krishna et al., 2020). We add two conservative systems that are challenging for automatic metrics: (i) a system that always copies the input and (ii) a content-preserving but nonsensical system that scrambles 5% of the input words.

Humans. We also add human-written simplifications using different instructions from ASSET (Alva-Manchego et al., 2020) and TurkCorpus (Xu et al., 2016), two widely used evaluation benchmarks for sentence simplification, and an auto-aligned one from the SimpleWiki (Kauchak, 2013).

D Implementation Details for Simplification Systems

D.1 T5 Setup

We use the Hugging Face Transformers library[9]. We fine-tune T5-base, T5-large, and T5-3B on 4 A40 GPUs of a total batch size of 64 for 20 epochs (10.4K steps), 16 epochs, and 8 epochs, respectively. We use a learning rate of 3e-4. We save checkpoints every 5K steps and select the best one by performing a manual inspection on a set of 60 simplifications from the development set, resulting in 80K, 60K, and 25K steps for T5-base, T5-large, and T5-3B respectively, which are in a similar range to FLAN-T5 (Chung et al., 2022). We use AdamW (Loshchilov and Hutter, 2017) as the optimizer.

D.2 GPT-3.5 Setup

Hyperparameters. We use the text-davinci-003 GPT-3.5 model from OpenAI API[10]. To generate simplification, we use the following hyperparameters: temperature=1 and top-p=0.95.

Prompts. We use the instruction from ASSET (Alva-Manchego et al., 2021) to prompt GPT-3.5.

Zero-shot setting: *Please rewrite the following complex sentence in order to make it easier to understand by non-native speakers of English. You can do so by replacing complex words with simpler synonyms (i.e. paraphrasing), deleting unimportant information (i.e. compression), and/or splitting a long complex sentence into several simpler ones. The final simplified sentence needs to be grammatical, fluent, and retain the main ideas of its original counterpart without altering its meaning.*

Input: {input}
Output:

5-shot setting: *Please rewrite the following complex sentence in order to make it easier to understand by non-native speakers of English. You can do so by replacing complex words with simpler synonyms (i.e. paraphrasing), deleting unimportant information (i.e. compression), and/or splitting a long complex sentence into several simpler ones. The final simplified sentence needs to be grammatical, fluent, and retain the main ideas of its original counterpart without altering its meaning.*

Examples:
Input: {random sampled from ASSET}
Output: {random sampled human reference of the Input from ASSET}

Input: {random sampled from ASSET}
Output: {random sampled human reference of the Input from ASSET}

Input: {random sampled from ASSET}
Output: {random sampled human reference of the Input from ASSET}

Input: {random sampled from ASSET}
Output: {random sampled human reference of the Input from ASSET}

Input: {random sampled from ASSET}
Output: {random sampled human reference of the Input from ASSET}

Please rewrite the following complex sentence in order to make it easier to understand by non-native speakers of English. You can do so by

[9] https://huggingface.co/
[10] https://beta.openai.com/

replacing complex words with simpler synonyms (i.e. paraphrasing), deleting unimportant information (i.e. compression), and/or splitting a long complex sentence into several simpler ones. The final simplified sentence needs to be grammatical, fluent, and retain the main ideas of its original counterpart without altering its meaning.

Input: {input}
Output:

E LENS metric - Additional Results

	SIMPEVAL$_{2022}$			WIKI-DA			NEWSELA-LIKERT		
	τ_{para}	τ_{spl}	τ_{all}	**Fluency**	**Meaning**	**Simplicity**	**Fluency**	**Meaning**	**Simplicity**
FKGL	-0.556	-0.31	-0.356	0.054	0.145	0.001	0.193	0.306	-0.051
BLEU	0.048	-0.054	-0.033	0.460	0.622	0.438	0.332	0.261	0.118
SARI	0.206	0.140	0.149	0.335	0.534	0.366	0.234	0.124	0.094
BERTScore	0.238	0.093	0.112	0.636	0.682	0.614	0.384	0.274	0.215
***Original* LENS**									
LENS$_{all}$	0.333	0.233	0.241	0.823	0.665	0.715	0.654	0.471	0.336
LENS$_{k=1}$	**0.460**	0.295	0.307	0.813	0.658	0.692	0.649	0.449	0.321
LENS$_{k=3}$	0.429	**0.333**	**0.331**	**0.843**	**0.684**	0.735	0.646	0.437	0.353
***Rescaled* LENS**									
LENS$_{all}$	0.333	0.233	0.241	0.816	0.662	0.733	**0.655**	**0.477**	0.343
LENS$_{k=1}$	**0.460**	0.295	0.307	0.796	0.647	0.721	0.633	0.444	0.328
LENS$_{k=3}$	0.429	**0.333**	**0.331**	0.807	0.668	**0.749**	0.624	0.428	**0.359**

Table 8: Metric evaluation results on SIMPEVAL$_{2022}$, WIKI-DA (Alva-Manchego et al., 2021), and NEWSELA-LIKERT (Maddela et al., 2021) human ratings datasets. We include the results for both original and rescaled versions of LENS. τ_{para}, τ_{spl}, and τ_{all} represent the Kendall Tau-like correlation for paraphrase-focused, split-focused, and all simplifications respectively. We report the Pearson correlation coefficients along three dimensions for WIKI-DA and NEWSELA-LIKERT. The best values are marked in **bold** and the second best values are underlined. As Kendall Tau measures pairwise rankings, we see the same results for the original and rescaled versions of LENS.

F Annotation Interface

F.1 Step - 1: System Output Categorization.

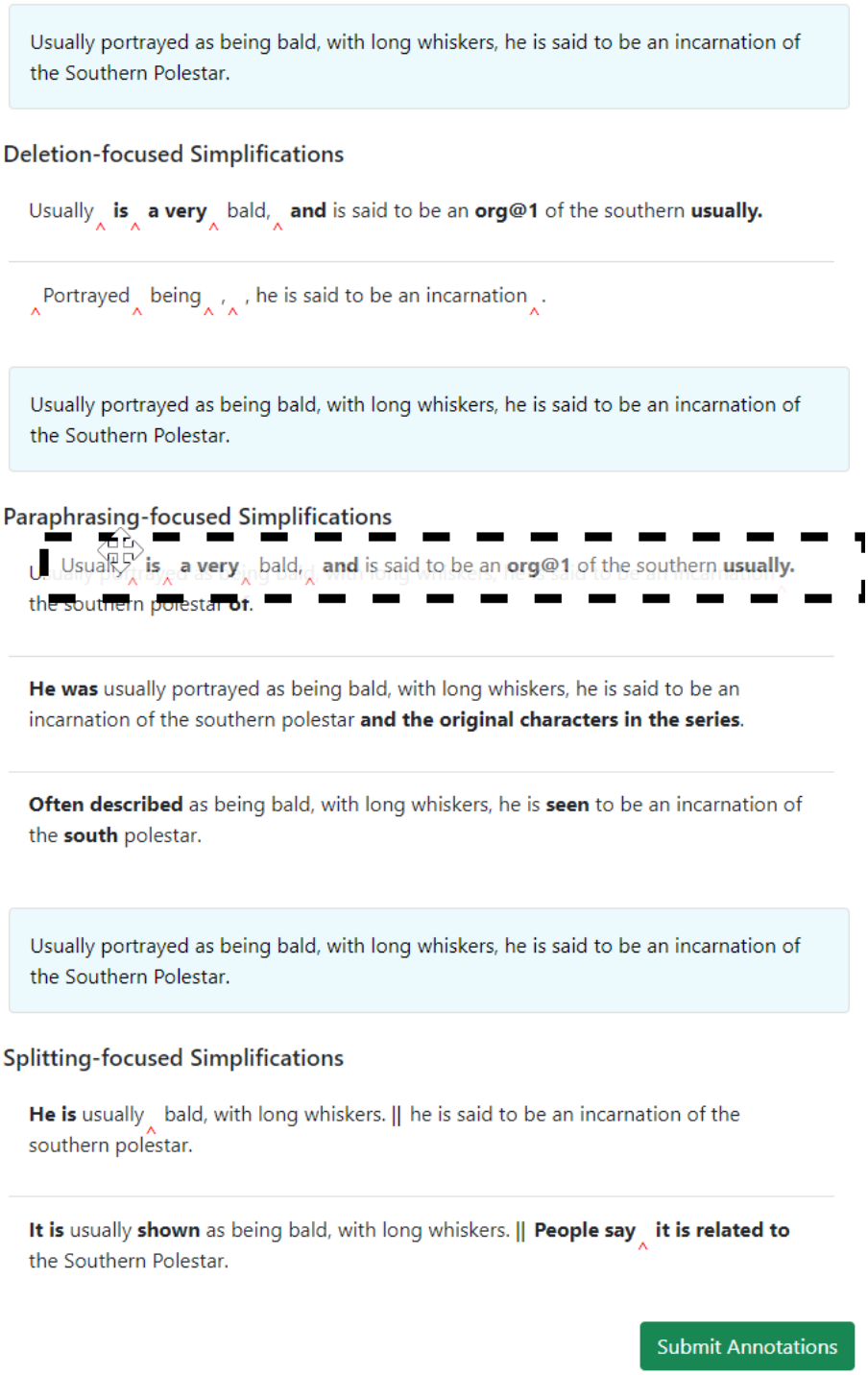


Figure 6: Annotation interface for categorizing system outputs. The outputs can be moved up and down or to other categories.

F.2 Step - 2: Highlighting System Edits.

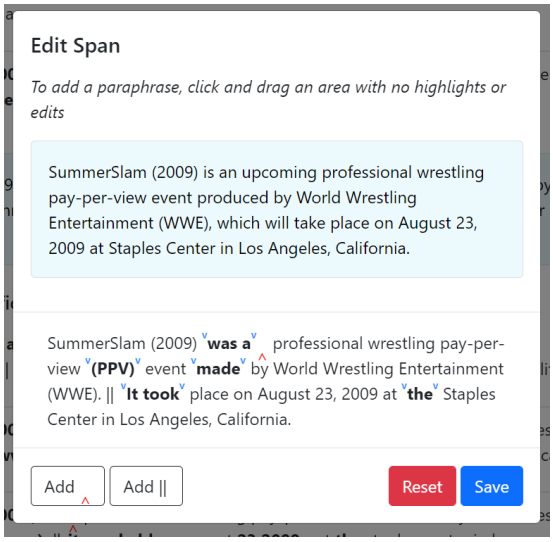


Figure 7: Span-fixing interface provided by annotators to fix the highlighted changes between the original and simplified sentences. Annotators could click and drag to modify the bounds of a paraphrase label or highlight text to add a paraphrase level. The deletion ("^") and split (" || ") can also be dragged and can be added using buttons on the bottom left of the modal.

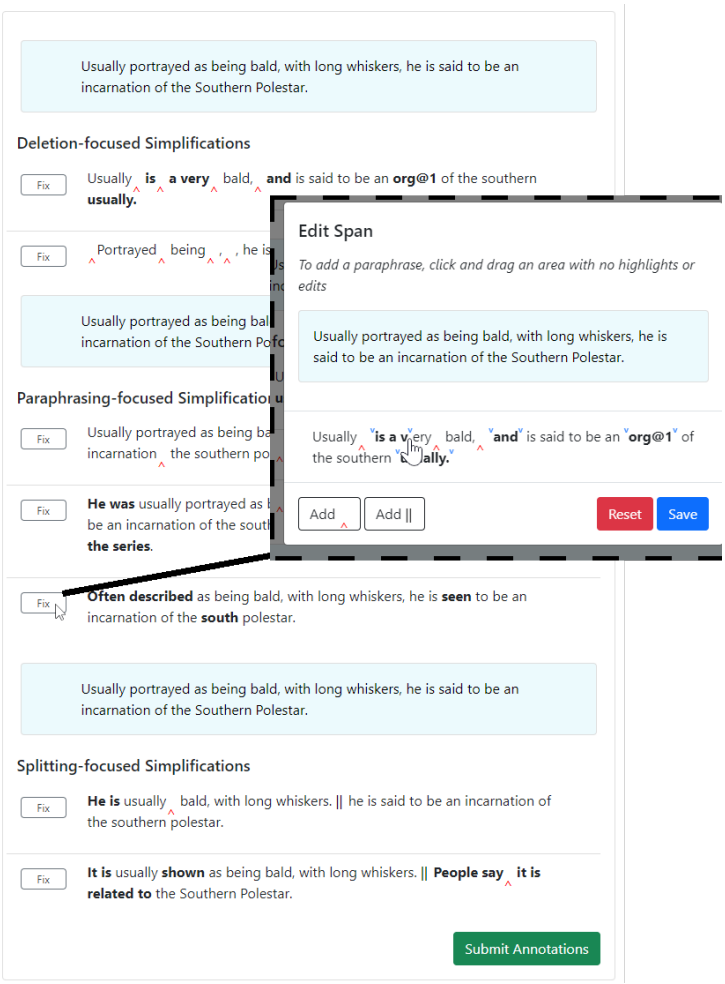


Figure 8: Annotation interface for span fixing. Clicking on "Fix" button opens up a pop up window.

F.3 Step - 3: Rating and Ranking System Outputs.

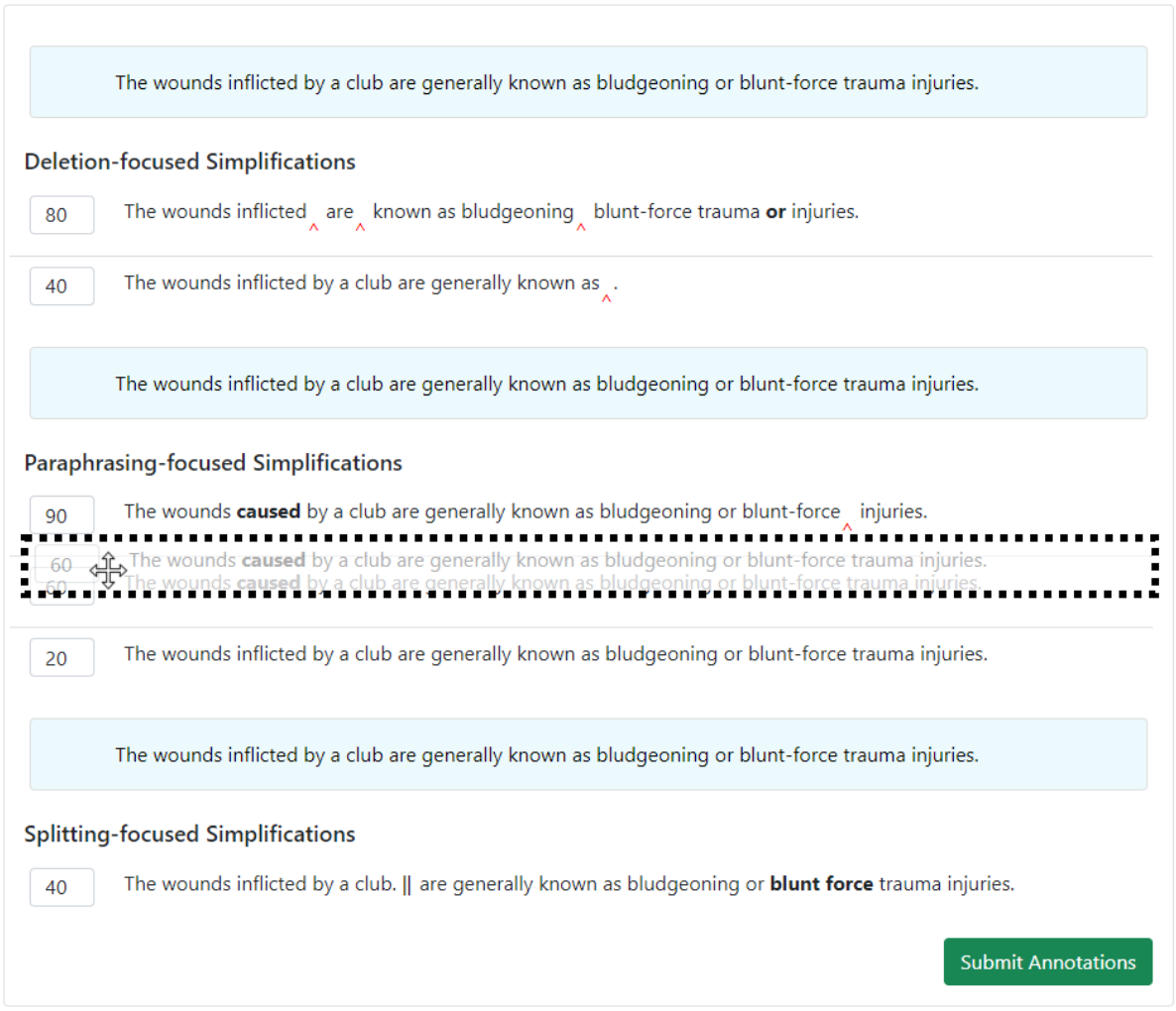
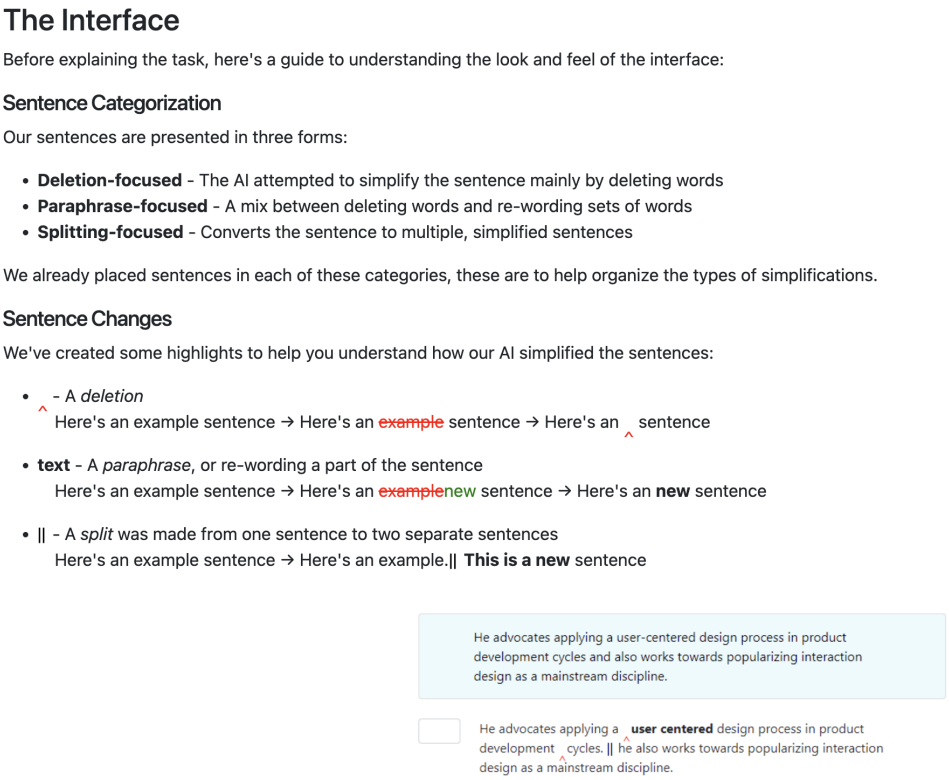


Figure 9: Sentence ranking and rating interface provided to annotators. The annotator can enter the ratings for each sentence and is able to re-order sentences by clicking and dragging their mouse.

F.4 Annotation Instructions



The Interface

Before explaining the task, here's a guide to understanding the look and feel of the interface:

Sentence Categorization

Our sentences are presented in three forms:

- **Deletion-focused** - The AI attempted to simplify the sentence mainly by deleting words
- **Paraphrase-focused** - A mix between deleting words and re-wording sets of words
- **Splitting-focused** - Converts the sentence to multiple, simplified sentences

We already placed sentences in each of these categories, these are to help organize the types of simplifications.

Sentence Changes

We've created some highlights to help you understand how our AI simplified the sentences:

- ^ - A *deletion*
  Here's an example sentence → Here's an ~~example~~ sentence → Here's an ^ sentence
- **text** - A *paraphrase*, or re-wording a part of the sentence
  Here's an example sentence → Here's an ~~example~~new sentence → Here's an **new** sentence
- || - A *split* was made from one sentence to two separate sentences
  Here's an example sentence → Here's an example.|| **This is a new** sentence

He advocates applying a user-centered design process in product development cycles and also works towards popularizing interaction design as a mainstream discipline.

He advocates applying a ^ **user centered** design process in product development ^ cycles. || he also works towards popularizing interaction design as a mainstream discipline.

Explanation: *Notice the paraphrases, deletions and sentence split is highlighted from the old to the new sentence.*

Figure 10: Instructions for the overall task.

Rating Sentences

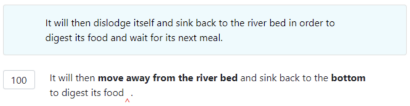
The other primary part of ths task is to **rate sentences** by how well they **simplify** the original sentence (according to our definition above).

One challenge is deciding how to score each sentence. Here's some general rules for scores:

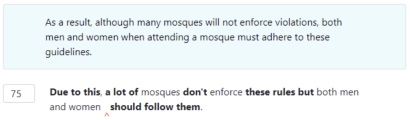
- **100** - Only when the sentence is fully simplified, entirely fluent and preserves the **core** meaning of the original sentence
- **75** - The sentence is somewhat simpler, mostly fluent and the meaning is close to the original sentence
- **50** - The sentence is simpler, somewhat fluent and the meaning is similar to the original sentence
- **25** - The sentence is equivalently simple, still has some fluency but the meaning is lost
- **0** - The sentence is completely unreadable

Most scores will lie somewhere in this range, feel free to give specific scores.

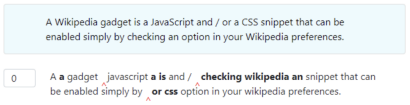
Examples:



Explanation: *Notice how (1), the sentence reads completely fluently and (2) the orignal meaning of the sentence ("it" getting unstuck and moving to the bottom)*



Explanation: *Although the sentence has a error in its fluency, it still has the same meaning as the orignal sentence*



Explanation: *The sentence is completely unreadable*

Figure 11: Instructions for ranking and rating.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations section

☑ A2. Did you discuss any potential risks of your work?
Limitations section Ethics Statement section

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

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

B ☑ Did you use or create scientific artifacts?

Section 3

☑ B1. Did you cite the creators of artifacts you used?
Section 3, Appendix sections A - D.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Sections 3 Ethics Statement section

☑ 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)?
Sections 3 Ethics Statement section

☑ 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?
Ethics Statement section

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Sections 3 Ethics Statement section

☑ 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.
Section 3

C ☑ Did you run computational experiments?

Section 4

☑ 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 Sections A - D

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?
Section 3 Appendix Sections A - D

☑ 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 4

☑ 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.)?
Appendix Sections A - D

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Section 3

☑ 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.?
Appendix section F

☑ 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)?
Sections 3 Ethics Statement section

☑ 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?
Ethics Statement section

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No. This study does not meet the definition of "human subjects" research by our institutional IRB review board, as we did not: "(1) Interact with, intervene with, or obtain/access private, identifiable information or data about, a living individual (includes online surveys) or (2) Conduct research on a drug, biologic, or medical device". The annotations are linguistic judgements provided by hired in-house employees.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Ethics Statement section