Meta-Evaluation of Sentence Simplification Metrics

Noof Alfear[*†], Dimitar Kazakov[*], Hend Al-Khalifa[†]
[*]Deptartment of Computer Science, University of York
York, UK
naa542, dimitar.kazakov@york.ac.uk

[†]Information Technology Department, King Saud University
Riyadh, Saudi Arabia
nalfear, hendk@ksu.edu.sa

Abstract

Automatic Text Simplification (ATS) is one of the major Natural Language Processing (NLP) tasks, which aims to help people understand text that is above their reading abilities and comprehension. ATS models reconstruct the text into a simpler format by deletion, substitution, addition or splitting, while preserving the original meaning and maintaining correct grammar. Simplified sentences are usually evaluated by human experts based on three main factors: simplicity, adequacy and fluency or by calculating automatic evaluation metrics. In this paper, we conduct a meta-evaluation of reference-based automatic metrics for English sentence simplification using high-quality, human-annotated dataset, NEWSELA-LIKERT. We study the behavior of several evaluation metrics at sentence level across four different sentence simplification models. All the models were trained on the NEWSELA-AUTO dataset. The correlation between the metrics' scores and human judgements was analyzed and the results used to recommend the most appropriate metrics for this task.

Keywords: simplification, evaluation, SARI, BLEU, BERTScore, LENS, NEWSELA

1. Introduction

Automatic Text Simplification (ATS) is a Natural Language Processing (NLP) task that aims to transform complex text to a simpler version of itself, while preserving correct grammar and the original meaning. Complex text is defined as text with a complex syntactic structure or difficult vocabulary. However, the simplicity level may vary depending on the goal behind the simplification process: either to be easier for humans to read and understand or for other NLP tasks to process. Children or second language learners or people with low literacy or cognitive impairments such as dyslexia, aphasia, autism or down syndrome face difficulties in understanding complex text. So, for these cases simplified text can enhance the reading comprehension. On the other side, NLP tasks like parsing, Machine Translation and information retrieval will have better performance when the input text is formulated in a clear and easy to understand structure (Siddharthan, 2014). ATS generated text can be evaluated either by language experts or scored automatically by calculating some well-defined metrics.

In this paper, we had an in-depth analysis of the relationship between human evaluation and reference-based automatic evaluation metrics across different state-of-the-art English sentence simplification models to define which metrics we can rely on when evaluating newly developed simplification models. Based on our findings, we recommend using LENS (Maddela et al., 2023), BERTScore$_{Precision}$, BERTScore$_{Recall}$ and BERTScore$_{F1}$ (Zhang* et al., 2020) when evaluating sentence simplification models. Although BLEU (Papineni et al., 2002) and SARI (Xu et al., 2016) are not correlated with human judgement, we recommend to continue reporting them for comparison with previously published state-of-the-art approaches.

2. Background

A major challenge in this research area that limits its development is the evaluation aspect and how accurately we can measure the quality and adequacy of simplification models' outcomes (Stajner et al., 2016). The simplified text should be evaluated based on three aspects: the simplicity level, grammar correctness (fluency) and original meaning preservation (adequacy). Researchers have followed various approaches in evaluating the outcomes of a simplification model. The most common and reliable approach depends on human experts to evaluate the simplified text on a 1-5 Likert scale taking into their accounts the three aspects mentioned earlier. However, it is hard to compare the output quality of different simplification models if they were evaluated by different experts. People are subjective and they have different opinions and views. Moreover, this process is time and cost consuming and alternative methods are needed

(Alva-Manchego et al., 2020b).

Another approach is depending on automatic evaluation metrics. Readability metrics are applied to measure the simplicity level of the text compared to the original, such as Flesch Kincaid, FOG index and many others. Those metrics are not sufficient because they rely on shallow features of the text like average sentence length or average number of syllables per word, etc. and ignore the simplicity, the grammar and semantic adequacy of the outcomes (Alva-Manchego et al., 2020b). Other researchers used the machine translation evaluation metric: BLEU, which compares the n-grams of the simplified text with n-grams of other references in the dataset. References are sentences simplified multiple times by human editors. However, BLEU is not entirely suitable because it penalizes for operations that are common in Text Simplification such as word deletion, insertion and reordering. Nevertheless, researchers continue using BLEU because they found that it sometimes correlates with human judgement scores of adequacy and fluency, even if not with simplicity (Xu et al., 2016). In 2016, Xu et al. designed the first metrics specifically for evaluating simplified text: FK-BLEU and SARI. FK-BLEU combines the paraphrase generation metric, iBLEU which is an extension to the BLEU metric with the readability metric, Flesch Kincaid Index. FKBLEU is suitable for evaluating simplification by paraphrasing rather than deletion or splitting, which does not cover different text simplification models. On the other hand, SARI compares the system output against other references and the input sentence. It rewards addition operations that occur both in the output and any of the references and words kept or deleted by both system output and any of the references. In 2018, Sulem et al. proposed SAMSA, which was the first automatic metric to quantify the structural aspects of simplified text not only lexicons as previous measures. This feature was achieved by assessing sentence splitting correctness compared to the input. BERTScore was also proposed by Zhang* et al. in 2020 to evaluate Text Generation tasks using BERT. It computes a similarity score for each token in the output with each token in the reference using contextual embeddings. Moreover, BERTScore computes precision, recall and F1 measure, which can be useful for evaluating different language generation tasks. The recent published Text Simplification evaluation metric is LENS which is a Learnable Evaluation Metric for Text Simplification trained on SimpEval corpus. This corpus contained SimpEval-past, which has 12K human ratings on 2.4K simplifications of 24 past systems and SimpEval-2022, which consists of over 1K human ratings of 360 simplifications including GPT-3.5 generated text (Maddela et al., 2023).

To our knowledge, the first meta-evaluation study of Sentence Simplification automatic evaluation metrics was conducted by Alva-Manchego et al. in 2021. However, this study includes a recently published metric for text simplification: LENS (Maddela et al., 2023), and the dataset used, along with Human Judgements on Simplicity (Maddela et al., 2021) differs from the previous study. Moreover, we studied the variation of both linear and non-linear measures of correlation, namely Pearson, Spearman (Schober et al., 2018) and Kendall's Tau (Puka, 2011), with respect to four different simplification models. All models were trained on the same high-quality dataset, NEWSELA-AUTO (Jiang et al., 2020). On the other hand, our study aligns with the findings of Maddela et al.. However, a distinction arises regarding the NEWSELA-LIKERT dataset. While Maddela et al. exclusively reported the linear correlation, Pearson, we provided both linear and non-linear correlations along with the significance levels. This choice was made because it is well-known that for ordinal data, the other two non-linear correlations, Spearman and Kendall's Tau, are more appropriate measures of correlation. Furthermore, our analysis of BERTScore included examination of precision, recall, and F1 values, utilizing the best-performing model as reported by Zhang* et al. In contrast, the previous study only reported precision values.

3. Meta-Evaluation of Automatic Evaluation Metrics

This study analyzes the relationship between the automatic evaluation metrics and human judgement for Sentence Simplification task across multiple simplification models. The goal is to find whether we can rely on those automatic evaluation metrics when evaluating sentence simplification models rather than depending on humans in the future. Moreover, this meta-evaluation inspects if metrics' correlations are affected by the type of the model that generated the simplifications considering that all of them trained on the same high quality parallel dataset (NEWSELA-AUTO).

The focus was on recently published reference-based metrics, specifically BERTScore and LENS along with traditional metrics commonly used for Sentence Simplification task: BLEU and SARI. SAMSA was not included in this study because its main premise is that a structurally correct simplification consists of each sentence containing a single event from the input, which does not align with the typical primary goal of general simplification (Sulem et al., 2018). General simplification involves various operations such as lexical or phrase substitution, splitting, paraphrasing, addition, and deletion, while SAMSA specifically address correct splitting. To evaluate the automatic metrics, we computed the correlations between metrics' scores and human judgements via Pearson, Spearman and kendall's Tau for each metric. Since Pearson only detects linear relations and is sensitive to outliers, rank correlations: Spearman and kendall's Tau helped us overcome this and can detect monotonic non-linear relationship. Furthermore, we followed (Alva-Manchego et al., 2021) in performing Williams significance tests (Williams, 1959) to detect if the correlation between two variables is statistically significant or not.

3.1. Dataset with Human Judgements on Simplicity (NEWSELA-LIKERT)

The meta-evaluation study was conducted on NEWSELA-LIKERT (Maddela et al., 2021). This dataset was created to evaluate the performance of Controllable Text Simplification model with Explicit Paraphrasing. It has human evaluation of the overall simplification quality of 100 random sentences from the NEWSELA-AUTO test set. The simplified sentences were generated by Maddela et al. model and three other state-of-the-art previous models: HYBRID (Narayan and Gardent, 2014), EDITNTS (Dong et al., 2019) and TRANSFORMER (Jiang et al., 2020), where all the models were trained on the NEWSELA-AUTO dataset. Each simplified sentence fluency, adequacy and simplicity was rated on a 5-point Likert scale by five Amazon Mechanical Turk workers, where 5 is the best and 1 is the worst. The ratings were averaged as the human ratings are fairly consistent, with very few outliers. As mentioned earlier, fluency assesses whether the output maintains correct grammar, adequacy verifies if the output preserves the original meaning of the input sentence and simplicity evaluates if the output is simpler than the input in terms of both lexicon and syntax.

There were three other datasets with human judgements: SIMPEVAL$_{PAST}$, SIMPEVAL$_{2022}$, and WIKI-DA. SIMPEVAL$_{PAST}$ contains 12K human ratings on 2.4K simplifications from 24 systems on sentences from TurkCorpus (Xu et al., 2016). This dataset was employed to train LENS. SIMPEVAL$_{2022}$ consists of 1,080 human ratings on 360 simplifications from both humans and state-of-the-art models, including GPT-3.5. This dataset was utilized to evaluate LENS and other simplification metrics (Maddela et al., 2023). WIKI-DA released by Alva-Manchego et al. in 2021 with 0-100 continuous scale ratings on fluency, adequacy, and simplicity for 600 simplifications across six systems. While SIMPEVAL$_{PAST}$, SIMPEVAL$_{2022}$, and WIKI-DA are derived from Wikipedia, NEWSELA-LIKERT is derived from news articles in Newsela. Those three datasets were excluded from the study because SIMPEVAL$_{PAST}$ and SIMPEVAL$_{2022}$ were involved in the development of LENS metric. On the other side WIKI-DA was the heart of the first Meta-Evaluation of Automatic Evaluation Metrics conducted by Alva-Manchego et al.

3.2. Methodology

At the beginning, non-referenced basic readability metrics were applied on NEWSELA-LIKERT dataset to get sense of how complex/simple sentences are evaluated across multiple sources and the results are shown in Table 1. Obviously, most of those readability metrics do not show any significant difference in the values across multiple sources, which means those metrics are not suitable to measure the simplicity of the sentences. Except the fog index, which might correctly detect the complexity level of the input sentence. From the table, the fog index for the input complex sentences has a score of 7.5, which means a seventh grader would be able to read and understand the sentences. On the other hand, Maddela et al. output sentences has the lowest score 4.78, that means a fourth grader can understand those sentences.

However, this study focuses on reference-based metrics BLEU, SARI, BERTScore and LENS to measure simplified sentences and investigate their correlations with simplicity, adequacy and fluency scores from human judgements. All the metrics were calculated at the sentence-level. For BLEU and SARI, the implementations provided by EASSE (Alva-Manchego et al., 2019) was applied. On the other side, for BERTScore both the implementation with RoBERTa as the default model and also microsoft/deberta-xlarge-mnli model (He et al., 2021) based on the authors recommendation (Zhang* et al., 2020) were used. The later was reported here because it had the best correlation with human evaluation. For LENS we used the implementation provided by Maddela et al. using RoBERTa with LENS(k=3).

After that, the three different correlation types: Pearson, Spearman and Kendall's Tau were calculated for each metric with human judgement different aspects' scores across the four simplification models: HYBRID (Narayan and Gardent, 2014), EDITNTS (Dong et al., 2019), Jiang et al. and Maddela et al. as shown in Tables 2, 3, 4 and 5 respectively. Multiple correlations were applied to detect any kind of relationship whether it was linear or non-linear.

	Complex	Reference	HYBRID	Maddela et al. (2021)	TRANSFORMER	EDITNTS
flesch reading ease	**89.89**	**89.28**	92.93	95.67	**86.91**	**87.82**
flesch kincaid grade	**4.5**	**4.7**	3.3	2.3	3.6	3.2
fog index	**7.5**	6.64	5.64	4.78	5.18	5.28
difficult words	**24**	16	16	16	23	18
syllable count	240	291	184	147	159	146
lexicon count	197	253	158	123	127	116

Table 1: Basic Readability Statistics - Highest in difficulty are marked in **bold**, while the least are underlined.

4. Results

The resulted correlation between the automatic evaluation metrics (BLEU, SARI, BERTScore$_{Precision}$, BERTScore$_{Recall}$, BERTScore$_{F1}$ and LENS) and the human judgement evaluation based on simplicity, adequacy and fluency are presented in Table 2, Table 3, Table 4 and Table 5 for the state-of-the-art simplification models: HYBRID (Narayan and Gardent, 2014), EDITNTS (Dong et al., 2019), TRANSFORMER (Jiang et al., 2020) and Maddela et al. (2021) respectively. Each table shows the three different correlations Pearson, Spearman and Kendall's Tau between each pair of automatic metrics and human judgements' aspects. All significant results with p < 0.05 are boldfaced.

5. Discussion

From the previous section and based on the interpretation of correlation coefficient values by Corder and Foreman, we can analyze the correlations across the multiple approaches. **LENS** shows moderate significant correlation with simplicity and medium to strong with Fluency for all the four different approaches. Also, it has significant moderate correlation with adequacy among all the models except TRANSFORMER.

Across the four different approaches three of them have an additional control layer over the simplification of the sentence either by splitting or deletion with the paraphrasing, while TRANSFORMER is the only vanilla simplification model. TRANSFORMER shows that most of the automatic metrics correlated with some aspects of human judgement. **BLEU**, **BERTScore**$_{Precision}$, **BERTScore**$_{Recall}$ and **BERTScore**$_{F1}$ have a medium significant correlation with Adequacy and Fluency. While **SARI** has absolute moderate significant correlation with Simplicity and Adequacy and **BERTScore**$_{Recall}$ has absolute moderate significant correlation with Simplicity. On the other side in Maddela et al., **BERTScore**$_{Precision}$, **BERTScore**$_{Recall}$ and **BERTScore**$_{F1}$ have medium significant correlation with Fluency. While **BERTScore**$_{Precision}$ and SARI have moderate significant correlation with Adequacy. In EDITNTS and HYBRID models, **BERTScore**$_{Precision}$, **BERTScore**$_{Recall}$ and **BERTScore**$_{F1}$ have moderate significant correlation with Fluency. Where **BERTScore**$_{Recall}$ correlates also with Adequacy. On the other side, **SARI** has significant medium correlation with Fluency and **BLEU** with Adequacy in the HYBRID model.

When analyzing the results and by looking at Tables 2, 3, 4 and 5, we can realize that some of Pearson correlations' values are significant, while Spearman and Kendall's Tau are not. Figures 1 and 2 show examples of this special case and how the data are distributed. Figure 1 shows a scatter plot of SARI scores VS Fluency for EDITNTS model. Where Pearson= **0.208836**, Spearman= 0.117007 and Kendall's Tau= 0.0781688, there is an outlier that affected the value of Pearson correlation. Another example is shown in Figure 2 for HYBRID model between BLEU and Fluency. Pearson = **0.221159**, Spearman = 0.151132 and Kendall's Tau = 0.100834. Again, the outlier affected Pearson correlation, while Spearman and Kendall's Tau show no significant correlation.

Overall, we can conclude that LENS has medium to strong correlation with all the human evaluation aspects: Simplicity, Adequacy and Fluency. While BERTScore$_{Precision}$, BERTScore$_{Recall}$ and BERTScore$_{F1}$ have medium correlation with Fluency and BERTScore$_{Recall}$ has moderate correlation with Adequacy. Transformer is the only model that has correlation of both BLEU with Fluency and SARI with Simplicity.

	Simplicity			Adequacy			Fluency		
	Pearson	**Spearman**	**kendall's Tau**	**Pearson**	**Spearman**	**kendall's Tau**	**Pearson**	**Spearman**	**kendall's Tau**
BLEU	-0.042704	-0.080227	-0.061295	**0.305895**	**0.247185**	**0.174170**	**0.221159**	0.151132	0.100834
SARI	0.098624	0.028061	0.021687	**0.281641**	0.171048	0.113217	**0.324004**	**0.220126**	**0.152129**
BERTScore$_P$	0.143842	0.068493	0.050753	0.140552	0.076047	0.049008	**0.276201**	**0.199609**	**0.143602**
BERTScore$_R$	0.085582	0.013768	0.016634	**0.364709**	**0.263457**	**0.186610**	**0.412355**	**0.323510**	**0.230593**
BERTScore$_{F1}$	0.123921	0.047653	0.035162	**0.274748**	0.181527	0.12322	**0.371118**	**0.286157**	**0.204717**
LENS	**0.284418**	**0.236678**	**0.176231**	**0.308810**	**0.265191**	**0.191197**	**0.553059**	**0.498933**	**0.365820**

Table 2: Correlation-HYBRID simplification model (significant correlations with $p<0.05$ are boldfaced)

	Simplicity			Adequacy			Fluency		
	Pearson	**Spearman**	**kendall's Tau**	**Pearson**	**Spearman**	**kendall's Tau**	**Pearson**	**Spearman**	**kendall's Tau**
BLEU	-0.177190	-0.140434	-0.103769	0.042482	-0.019924	-0.025588	0.056722	-0.016503	-0.010145
SARI	0.184546	0.156650	0.116005	0.174147	0.130178	0.080087	**0.208836**	0.117007	0.078169
BERTScore$_P$	0.138181	0.083496	0.055118	0.142600	0.123742	0.087933	0.118100	**0.07011**	0.081614
BERTScore$_R$	-0.011886	-0.011435	-0.005126	**0.234410**	**0.226968**	**0.159175**	0.109790	**0.323510**	0.101147
BERTScore$_{F1}$	0.077024	0.052734	0.035033	0.197942	0.193782	0.134521	0.125228	**0.286157**	0.105195
LENS	**0.496585**	**0.505682**	**0.364454**	**0.452231**	**0.344530**	**0.239375**	**0.601925**	**0.453529**	**0.326975**

Table 3: Correlation-EDITNTS simplification model (significant correlations with $p<0.05$ are boldfaced)

	Simplicity			Adequacy			Fluency		
	Pearson	**Spearman**	**kendall's Tau**	**Pearson**	**Spearman**	**kendall's Tau**	**Pearson**	**Spearman**	**kendall's Tau**
BLEU	-0.067261	-0.075353	-0.051788	**0.378386**	**0.370779**	**0.262711**	0.177556	**0.246839**	**0.175007**
SARI	**-0.203313**	**-0.216973**	**-0.151497**	**0.275280**	**0.274311**	**0.191213**	0.139153	0.121658	0.087574
BERTScore$_P$	0.139345	0.131615	0.090240	**0.314171**	**0.307780**	**0.207120**	**0.228873**	**0.308736**	**0.187060**
BERTScore$_R$	**-0.231895**	**-0.231675**	**-0.162777**	**0.391481**	**0.411641**	**0.288427**	0.172463	**0.323510**	0.134079
BERTScore$_{F1}$	-0.056910	-0.059074	-0.046282	**0.380872**	**0.392788**	**0.265581**	**0.214277**	**0.286157**	**0.176668**
LENS	**0.288571**	**0.339714**	**0.23874**	0.156929	0.109114	0.074516	**0.429411**	**0.423822**	**0.306160**

Table 4: Correlation-TRANSFORMER simplification model (significant correlations with $p<0.05$ are boldfaced)

	Simplicity			Adequacy			Fluency		
	Pearson	**Spearman**	**kendall's Tau**	**Pearson**	**Spearman**	**kendall's Tau**	**Pearson**	**Spearman**	**kendall's Tau**
BLEU	-0.116446	-0.112289	-0.079251	-0.119405	-0.079144	-0.061568	0.130129	0.194735	0.135584
SARI	-0.007717	0.019843	0.008117	**0.259365**	**0.304577**	**0.219893**	0.102994	0.170013	0.111771
BERTScore$_P$	0.138181	0.083496	0.055118	0.142600	0.123742	0.087933	0.117997	**0.169106**	0.081614
BERTScore$_R$	-0.011885	-0.011435	-0.005126	**0.234410**	**0.226968**	**0.159175**	0.109790	**0.323510**	0.101147
BERTScore$_{F1}$	0.077024	0.052734	0.035033	0.197942	0.193782	0.134521	0.125228	**0.286157**	0.105195
LENS	**0.284043**	**0.255757**	**0.176624**	**0.347847**	**0.367430**	**0.265194**	**0.405085**	**0.432066**	**0.310846**

Table 5: Correlation-Mounica simplification model (significant correlations with $p<0.05$ are boldfaced)

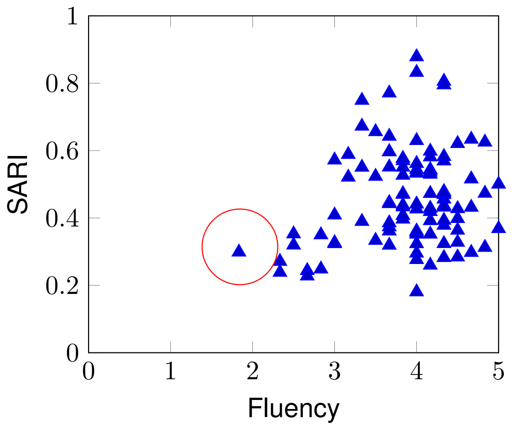

Figure 1: EDITNTS Fluency vs SARI

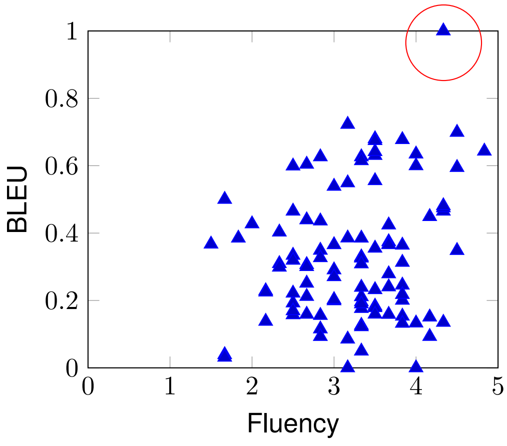

Figure 2: HYBRID Fluency vs BLEU

From this analysis and based on our findings, LENS is recommended at the first place to measure the three different aspects of simplified sentences: Simplicity, Adequacy and Fluency. While BERTScore$_{Precision}$, BERTScore$_{Recall}$ and BERTScore$_{F1}$ are precise measures of simplified sentences' Fluency. Moreover, BERTScore$_{Recall}$ can be used to support the meaning preservation factor. On the other side, we recommend BLEU and SARI continue to be reported to help researchers compare the performance of new models with earlier published results.

6. Conclusion

In this paper, the degree of which a reference-based automatic evaluation metric can measure the quality of a sentence simplification model was studied across multiple simplification models. The dataset; NEWSELA-LIKERT; consists of 400 automatic simplifications generated by four state-of-the-art systems, three of which are based on modern neural sequence-to-sequence architectures.

Our meta-evaluation study concludes that LENS is one of the best reference-based metrics to use with sentence simplification evaluation. It was able to measure the three different aspects: Adequacy, Fluency and Simplicity. The study also recommends using BERTScore$_{Precision}$, BERTScore$_{Recall}$ and BERTScore$_{F1}$ to measure simplified sentence Fluency. Furthermore, BERTScore$_{Recall}$ can be applied to measure the meaning preservation of the simplified sentence compared to the complex input sentence. Finally, although SARI and BLEU did not show strong correlation with any of the aspects, we still recommend reporting them in new publication to help comparing them with previous published state-of-the-art models.

7. Acknowledgements

We express our gratitude to Newsela for sharing the data and to Maddela et al. (2021) for providing access to NEWSELA-LIKERT, the dataset with human judgement evaluations.

8. Bibliographical References

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

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

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020b. Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, 46(1):135–187.

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.

G.W. Corder and D.I. Foreman. 2009. *Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach*. Wiley.

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.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

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.

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

Mounica Maddela, 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.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Llukan Puka. 2011. *Kendall's Tau*, pages 713–715. Springer Berlin Heidelberg, Berlin, Heidelberg.

Patrick Schober, Christa Boer, and Lothar A Schwarte. 2018. Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia*, 126(5):1763–1768.

Advaith Siddharthan. 2014. A survey of research on text simplification. *ITL - International Journal of Applied Linguistics*, 165:259–298.

Sanja Stajner, Maja Popovic, Horacio Saggion, Lucia Specia, and Mark Fishel. 2016. Shared task on quality assessment for text simplification.

Elior Sulem, Omri Abend, and Ari Rappoport. 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)*, pages 685–696, New Orleans, Louisiana. Association for Computational Linguistics.

E.J. Williams. 1959. *Regression Analysis*. WILEY SERIES in PROBABILITY and STATISTICS: APPLIED PROBABILITY and STATIST ICS SECTION Series. Wiley.

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.

9. Language Resource References

Maddela, Mounica and Alva-Manchego, Fernando and Xu, Wei. 2021. *Controllable Text Simplification with Explicit Paraphrasing*. Association for Computational Linguistics. [link].

Newsela. *Newsela* | *Request Newsela Data*. [link].