Evaluating automatic text simplification performance on classic literature

Brandon Wilde
Montclair State University
wildeb1@montclair.edu

Melissa Lopez
Montclair State University
lopezm44@montclair.edu

Abstract

Automatic text simplification (ATS) aims to improve the accessibility of written language through simplified structure and vocabulary. Most parallel corpora for the training and evaluation of sentence-level simplification come from informational texts, namely Wikilarge and Newsela. In this paper, we evaluate the performance of state-of-the-art ATS systems ACCESS and MUSS on a corpus of 150 English sentences sampled from classic literature. We find results via automated evaluation consistent with the reported performance on the Wikilarge test set. We also find that BERTScore correlates better with human scores than other commonly-used ATS metrics, such as SARI. Finally, we discuss challenges that persist in ATS research despite promising results against human references.

1 Introduction

1.1 Motivation

The primary goal of text simplification is to reduce the complexity of a text while preserving its most salient semantic aspects. Such simplified texts are valuable for numerous populations, including second-language learners and those with low literacy levels, learning disabilities (e.g., autism, ADHD, and dyslexia), acquired language disorders (e.g., aphasia), and other cognitive impairments (Štajner, 2021).

Access to simplified texts may encourage more engagement with literature and textual education materials. Overall quality of life may also be enhanced by the availability of age- and ability-appropriate reading materials. These include information sources (e.g., official documents and news articles that may impact one's life), intellectually stimulating media, and leisure reading materials that facilitate entertainment and good conversation.

1.2 Automatic text simplification

While manual text simplification may include high-level conceptual simplification at the document level, automatic text simplification (ATS) has traditionally focused on reducing syntactic complexity (i.e., using simpler sentence structure) and lexical complexity (i.e., using simpler words) at the sentence level.

These simplification strategies are largely ascribable to the narrower edit operations of split, merge, reorder, insert, delete (all reducing syntactic complexity) and transform/substitute (reducing lexical complexity) (Brunato et al., 2022).

ATS models require parallel corpora of complex and simplified sentence pairs for evaluation (and training, for supervised approaches). Due to the limited availability of corpora, state-of-the-art ATS systems have primarily focused texts from informational domains (e.g., news and reference sources) (Brunato et al., 2022). There is currently limited investigation into how state-of-the-art models perform on other domains, such as classic literature.

1.3 Our contribution

In order to begin investigating the prospect of ATS in non-informational domains, we build a small aligned simplification corpus from classic novels. Published state-of-the-art ATS systems are then re-assembled and evaluated on our corpus. We examine the extent to which existing ATS systems can generalize to this type of media and probe the results for areas of weakness that can be targeted in future work. In addition, we examine the relationship between automated metrics and human judgment on this corpus.

2 Related Work

2.1 ATS corpora

ATS is often considered a machine translation problem in which the translation is from more complex

to more simplified text within the same language. As a result, the training and evaluation of ATS models requires parallel corpora, i.e., corpora consisting of matched pairs of complex and simple sentences.

Several researchers have highlighted the difficulty of agreeing on a common goal and standard for text simplification (Štajner, 2021; Brunato et al., 2022). Siddharthan (2014) noted that since there is no native speaker of "Simplified English", the most appropriate human reference source is not evident and may vary according to the needs of a given population.

Due to the variable needs of different consumer populations, and perhaps also insufficient commercial incentive, few large corpora have been assembled for the training of ATS systems (Štajner, 2021) and most are from news sources (Brunato et al., 2022).

Two parallel English corpora have become the most widely used for research and development in ATS: Newsela and Wikilarge. Newsela is a proprietary collection of news articles and corresponding simplified versions, written by professional editors (Xu et al., 2015). Wikilarge, conversely, is a publicly available corpus comprising aligned sentences from English Wikipedia and Simple English Wikipedia (Zhang and Lapata, 2017).

Both of these corpora are often used for ATS model evaluation, but the breadth of their combined coverage is still narrow. Both data sets consist solely of nonfiction informational media, leaving other domains unevaluated. The Newsela corpus includes simplifications at 4 different reading levels, expanding its utility, but these 4 gradations do not adequately address all ATS needs either.

While other data sets have been compiled, they are not as large and most are similarly limited to informational texts (Štajner, 2021).

2.2 ATS models

Approaches to ATS have evolved in recent decades, particularly with the advent of machine learning. While the earliest approaches relied on statistical models to extract the most informative parts of a given text, most current methods employ deep learning models to generate new text in a sequence-to-sequence fashion (Sikka and Mago, 2020).

We concentrate our study on current ATS systems, of which there are two primary classes: supervised and unsupervised. These classes correspond to the training approach, specifically whether the systems rely on manually simplified texts as their teacher (supervised) or whether they are able to "self-teach" (unsupervised). Supervised models historically outperform their unsupervised counterparts, but the supervised systems' reliance on precious human-generated data leads many to pursue unsupervised approaches. We also predict that supervised systems will be constrained by the types and range of data they are trained on, allowing unsupervised systems to perform better in data-scarce domains.

ACCESS is a novel supervised sequence-to-sequence approach that is intended to confer more control to the user than traditional supervised models' mimicry offers (Martin et al., 2019). Although one could potentially constrain a model's output to conform to some desired simplification schema, ACCESS instead imposes soft constraints at the input. Specific well-defined strings ("control tokens") are prepended to each input, indicating the types and degrees of simplification for that particular example. During training, the control tokens reflect actual differences between the original and simplified sentences, then during inference the control tokens can be manipulated freely to guide the output toward the user's preferred form.

The four control tokens, and the operations they proxy, are relative text length (sentence compression), character-level text similarity (rewording/paraphrasing), relative middle word rank (lexical simplification), and relative dependency tree depth (syntactic simplification).

MUSS (Martin et al., 2020) is an unsupervised variation of the ACCESS approach. While retaining the same control token paradigm, MUSS incorporates a new method for mining simplification data from raw text. This creates an ideal opportunity (a minimal pair, one might say) for studying the differential outcomes of supervised and unsupervised ATS systems. Incidentally, both systems achieved state-of-the-art results at the time of their publishing (although the best MUSS evaluation scores were achieved by supplementing the model with human-annotated data). Model differences are summarized in Table 1.

2.3 ATS evaluation

Previous work in ATS typically reports performance on a variety of automated metrics as well as human evaluation.

Model	Base Model	Training Data
ACCESS	Transformer	Wikilarge
MUSS	Pretrained BART	Mined English paraphrases from CCNet
MUSS+	Pretrained BART	Mined English paraphrases from CCNET, Wikilarge

Table 1: Models used in this study and their distinguishing features. All models employ the control token stratagem.

2.3.1 Automated metrics for ATS

SARI (Xu et al., 2016) is a metric developed specifically for evaluating text simplification models and, as such, is considered first when ranking different models/outputs.

As stated in its name, the SARI method compares System output Against (multiple) References and against the Input sentence. Each SARI score considers the F1 scores of proper n-gram additions, deletions, and constants. The arithmetic mean of these F1 scores is then taken as the text's SARI score.

The Bilingual Evaluation Understudy (BLEU) method measures the proportion of n-grams that two texts have in common, and is a common metric for sequence-to-sequence NLP tasks, owing to its frequent correlation with grammaticality and meaning preservation (Papineni et al., 2002; Štajner et al., 2014). While it also serves this function in ATS, it should not be relied upon as the primary metric, because it has been observed to correlate inversely with sentence simplicity (Sulem et al., 2018).

More recently, BERTScore has been proposed for evaluating generative language models (Zhang et al., 2019). Using contextualized embeddings to represent the generated and reference sentences, one averages the token cosine similarity scores between the two while upweighting key (high idf score) tokens. BERTScorePrecision, in particular, has been advocated for as a strong ATS quality measure (Alva-Manchego et al., 2021).

Besides these reference-based metrics, a commonly reported sentence complexity metric is the Flesh-Kincaid Grade Level (FKGL) score (Kincaid et al., 1975), which generates a score based on sentence length (syntactic complexity) and word length (lexical complexity).

2.3.2 Human evaluation of ATS

Although automated evaluation is cheap, quick, and easily repeatable, the metrics can be challenging to interpret or influenced by the characteristics of the human references they are compared against (Van Der Lee et al., 2019). Therefore, human evaluation continues to be the gold standard for providing a more complete understanding of a system's performance.

There are two main approaches to human evaluation of ATS and natural language generation (NLG) tasks in general. The most common method is the intrinsic approach, in which human annotators explicitly evaluate the quality of a system's predictions (Van Der Lee et al., 2019; Celikyilmaz et al., 2020). Extrinsic approaches evaluate the effectiveness (Gatt and Krahmer, 2018) or the usability (Štajner, 2021) of the end result. In the case of ATS, researchers may compare reading comprehension scores after readings of complex and simplified versions of a text. Few ATS studies have used an extrinsic approach (Rello et al., 2013; Saggion et al., 2015; Orăsan et al., 2018). It can be particularly costly and time-consuming, requiring special expertise, recruitment from specific populations, and careful consideration of environmental factors (Štajner, 2021). We focus on intrinsic evaluation given its frequency in the literature and its feasibility for our study.

Despite the frequency of an intrinsic approach to human evaluation in ATS and NLG, there is a lack of consistency in definitions, methods, and reporting. Howcroft et al. (2020) found numerous terms across the literature for similar evaluation concepts (e.g., fluency, readability, etc.), a variety of evaluation methods (e.g., direct, relative, and post-edit-based assessments), and gaps in reporting about these characteristics. Recent ATS studies have demonstrated relative consistency in method: they ask annotators to directly rate simplified text on a rating scale (typically a five-point scale: 0-4) on three quality dimensions: *Fluency*, *Simplicity*, and *Adequacy* (Xu et al., 2016; Kumar et al., 2020; Martin et al., 2020).

Fluency and Adequacy are among the most common quality dimensions across NLG since they are relevant to a variety of NLG tasks (Celikyilmaz et al., 2020). Fluency refers to the grammatical and mechanical correctness of the target text, while ad-

equacy refers to meaning retention from the source text to the target. In addition, the Simplicity dimension is used in ATS evaluation to evaluate the relative simplicity of the predicted text to its source. Although there is some consistency in dimensions used in recent ATS work, the authors often include limited information about the definitions of the scales used and the reliability of these evaluations.

3 Method

We constructed a small corpus of sentences from classic literature, simplified each sentence using state-of-the-art ATS models, and evaluated performance using manual and automated metrics.

3.1 Corpus construction

To construct a parallel corpus of literary texts, we visited the Project Gutenberg (Project Gutenberg) website for freely available classic literature texts. We selected 22 books from which to collect example sentences. Sentence-level simplifications were then collected from available abridged versions of ten of the books (Laybourn, 2017), and sentences from the other twelve books were simplified by the project's authors. The aim of manual simplification was to support comprehension by an adult with a lower reading level. Simplification operations included sentence splitting, paraphrasing, reordering, and lexical simplification.

3.2 Automatic text simplification procedure

For each sentence, we collected a system prediction from three text simplification models developed by Facebook Research: ACCESS (trained on Wikilarge parallel corpus), MUSS (trained on paraphrases mined from the Common Crawl data set), and MUSS+ (trained on both the mined paraphrase data set and the Wikilarge corpus). Control token targets were initially set to the default settings in MUSS (Compression ratio: 0.9, Levenshtein ratio: 0.65, and Word Rank ratio: 0.75). After collecting quality estimation metrics on the human-simplified sentences using EASSE (i.e., compression ratio, Levenshtein similarity, and lexical complexity score), we adjusted the control tokens to approximate the characteristics of the human data set (Compression ratio: 0.75, Levenshtein ratio: 0.7, and Word Rank ratio: 0.95) and executed the models a second time.

3.3 Evaluation

We performed evaluations of the system outputs to compare systems, analyze errors, and explore the relationships between common automated metrics and human scores on our corpus.

3.3.1 Automated metrics

We used the EASSE toolkit (Alva-Manchego et al., 2019) to calculate automated metrics. EASSE is a Python package designed to facilitate evaluation of ATS systems on a number of measures. With EASSE, we computed metrics that compare system predictions to a human-simplified reference (BLEU, SARI, F1 by token, and BERTScore). We acknowledge that metrics such as BLEU and SARI benefit from multiple references and our single human reference may not provide the most robust evaluation. In addition to these metrics provided by EASSE, we calculated sentence-level cosine similarity between the predictions and corresponding reference sentences via the SentenceTransformers all-MiniLM-L6-v2 model (Reimers and Gurevych, 2019).

We also used EASSE to compute measures comparing the predictions to the source data (compression ratio, Levenshtein similarity, addition operation, number of sentence splits, and proportions of additions and deletions) and independent measures of complexity (FKGL, lexical complexity score). These measures were collected at both the sentence level and corpus level.

3.3.2 Human evaluation

In addition to automated metrics, we collected human evaluations to compare with automated metrics and support error analysis. We evaluated the system outputs on three dimensions: *Fluency*, *Simplicity*, and *Adequacy*. Each metric was judged on a scale of 0-4, with 4 referring to a more successful sentence. A total of 210 sentences (30 sentences each from the six model variations plus the human references) were annotated by the project's authors. Ten sentence groups (70 total sentences) were double-annotated to assess inter-annotator reliability.

In this evaluation, Fluency judgments related only to the target sentence (i.e., "How fluent/grammatical is this sentence (without comparing to the original)?"), while Simplicity and Adequacy compared the complex and simplified versions (i.e., "How much simpler is the target sentence than the original?" and "How well does the

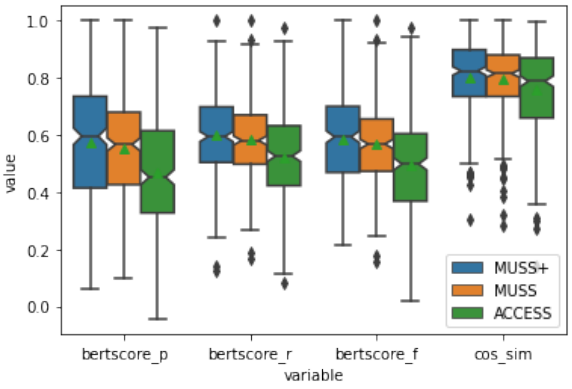

Figure 1: Automated metric score distributions by model (embedding-based metrics). This box plot shows the distributions for the models with human-matched control token targets.

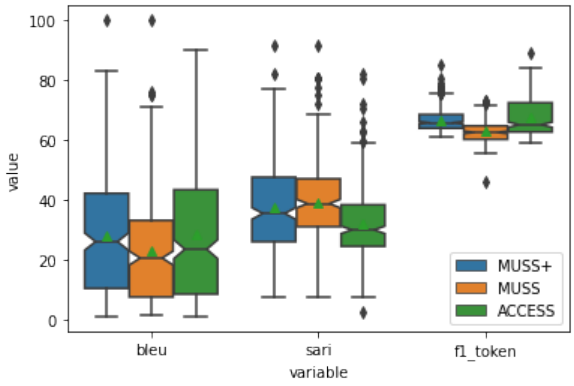

Figure 2: Automated metric score distributions by model (n-gram-based metrics). This box plot shows the distributions for the models with human-matched control token targets.

target sentence retain the meaning of the original?", respectively). Brief guidelines used for calibrating on the scale are provided in Appendix A (Table A1).

After annotation was completed, we computed sentence-level scores and overall means for each system and measured correlation between human means and automated metric means by system. We also measured correlation with sentence-level automated metrics with data from all systems.

4 Results

4.1 Automated evaluation results

The corpus-level mean scores across metrics are shown in Table 2. We refer to each system by the model name, and the control token parameters used, e.g., MUSS_default for the MUSS model trained on mined paraphrase data using default tokens, MUSS+_matched for the MUSS model trained on mined data and Wikilarge and collected using human-matched control tokens, etc. Some examples of sentences are provided in Appendix A (Table A2).

4.1.1 Reference-based metrics

MUSS+ performed best on automated metrics using embeddings. BERTScores are highest for the MUSS model using both Wikilarge and mined paraphrase data; the BERTScore for precision was higher for MUSS+_matched, whereas the BERTScore for recall was higher using MUSS+_default. This pattern is consistent across models and is likely due to the difference in compression ratio (0.9 by default and 0.75 to match humans). Sentence-level cosine similarity is highest for both MUSS model sets simplified with the default targets (0.80) and lowest for ACCESS_matched (0.75).

SARI scores range from 36.9 (ACCESS) to 43.6 (MUSS_default). F1 score was lowest for MUSS_matched at 58.9 and highest for MUSS+_default at 61.8. ACCESS_matched had the highest BLEU score at 37.5, while MUSS_default had the lowest. BLEU is the only reference-based metric where ACCESS performs best. Distributions in box plot form are shown in Figures 1 and 2.

4.1.2 Complexity measures

Complexity measures and other descriptive measures are presented in Table 3. All systems produced texts at a lower FKGL than (i.e., simpler than) the original (12.36). MUSS+_matched's FKGL (5.5) most closely approximated the human reference score (5.6). Most other systems had a FKGL score around 7-7.5, with ACCESS_default producing the least simple texts according to the FKGL measure. Conversely, the lexical complexity score for ACCESS_default was the lowest, suggesting that the FKGL score was higher due to longer sentences rather than more complex words. This is confirmed when noting that the ACCESS_default compression ratio is similar to close to 1. All of the human-matched runs resulted in more compressed (shorter) sentences than those using default control targets; MUSS_matched produced the shortest sentences. Overall, MUSS+_matched's results suggest that this version most closely approximates the complexity of the human references.

Version	SARI	BERTScore (Recall)	BERTScore (Precision)	BERTScore F1	BLEU	F1 by token	Cosine sim.
MUSS+ (m)	41.88	0.58	**0.59**	**0.59**	33.68	61.08	0.80
MUSS+ (d)	39.59	**0.62**	0.56	**0.59**	31.81	**61.83**	0.80
MUSS (m)	42.23	0.56	0.57	0.56	27.51	58.88	0.79
MUSS (d)	**43.63**	0.61	0.54	0.57	26.51	60.20	**0.80**
ACCESS (m)	36.94	0.51	0.50	0.51	**37.47**	60.66	0.75
ACCESS (d)	36.93	0.53	0.44	0.49	31.04	59.16	0.76

Table 2: Corpus-level reference-based metrics. Systems with the highest scores for each metric are presented in boldface. (m) means the control token targets were matched to reference characteristics, and (d) means the default settings were used.

Version	FKGL	Lexical complex.	Compress. ratio	Lev. sim.	Exact copies	Additions proportion	Deletions proportion	Sentence splits
Complex	12.36	8.73	-	-	-	-	-	-
Human	5.58	8.4	0.75	0.72	0.18	0.16	0.38	1.36
MUSS+ (m)	5.50	8.48	0.75	0.73	0.0	0.17	0.38	1.4
MUSS+ (d)	7.40	8.64	0.89	0.76	0.0	0.20	0.28	1.36
MUSS (m)	6.88	8.55	0.74	0.69	0.0	0.16	0.41	1.18
MUSS (d)	7.45	8.26	0.87	0.70	0.0	0.25	0.35	1.31
ACCESS (m)	7.14	8.48	0.77	0.81	0.01	0.08	0.29	1.19
ACCESS (d)	7.52	8.18	0.96	0.81	0.0	0.21	0.22	1.38

Table 3: Corpus-level descriptive measures. Average calculations of automated measures describing the resulting text in isolation (FKGL, Lexical complexity) and in relation to the the source text. (m) means the control token targets were matched to reference characteristics, and (d) means the default settings were used.

4.1.3 Other descriptive measures

With regard to the number of specific operations, ACCESS_matched deleted the smallest proportion of words (0.22), while MUSS_matched deleted the greatest proportion. Consistent with the results for FKGL and compression ratio, those matched to human parameters produced a greater number of deletions and fewer additions. Finally, the Levenshtein similarity (with reference to the original) is highest for ACCESS, followed by MUSS+ and MUSS, suggesting that MUSS predictions are the most different from the human references at the character level. MUSS+_matched produced the most sentence splits, with MUSS_matched producing the fewest. Again, MUSS+_matched performed most consistently with the human measures.

4.2 Human evaluation results

Average human evaluation scores are reported in Table 4, and distributions for the human-matched results are shown in Figure 3. Human-simplified texts received the highest Fluency scores on average, followed by MUSS+, MUSS, and ACCESS. Models run with target ratios matching the human simplifications in the corpus consistently received higher scores than their default counterparts.

Model	Fluency	Simplicity	Adequacy
Human	3.95	2.9	3.78
MUSS+ (m)	**3.8**	**3.2**	**3.28**
MUSS+ (d)	3.78	2.0	3.45
MUSS (m)	3.65	2.78	3.13
MUSS (d)	3.55	2.38	3.08
ACCESS (m)	2.28	1.68	2.13
ACCESS (d)	2.05	1.0	2.15

Table 4: Average human ratings by model. Mean scores across all human-evaluated data rated from 0-4 (with 4 being the highest). ATS system with the highest scores for each column are show in boldface. (m) means the control targets were matched to reference characteristics, and (d) means the default settings were used.

Human simplifications also received the highest scores for Adequacy, indicating that they preserved meaning most reliably. Scores follow a pattern similar to Fluency except that the control parameters didn't have the same effect on the final result.

In the evaluation of Simplicity,

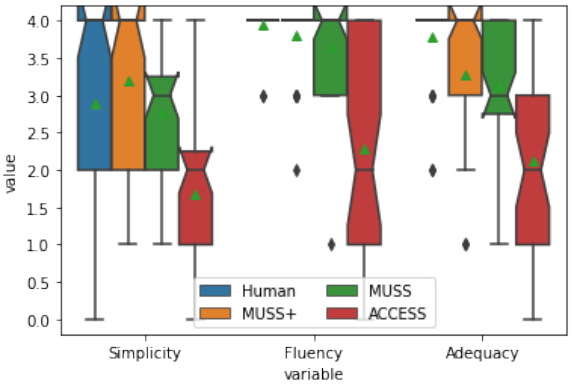

Figure 3: Human rating distributions by model. This box plot shows the distributions of human ratings for the models with human-matched control targets.

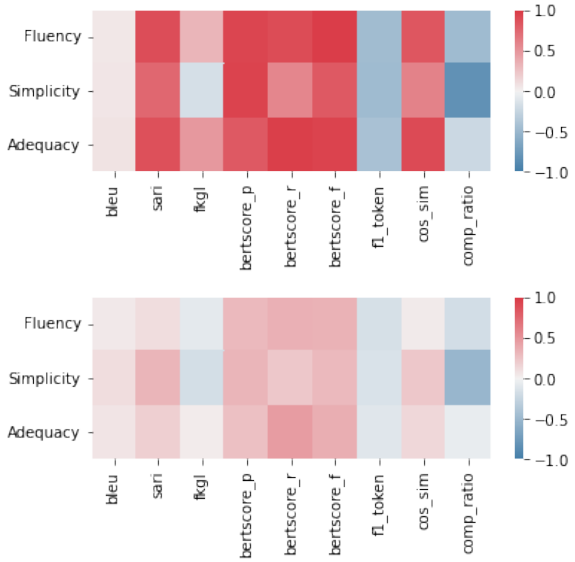

Figure 4: Heat maps for Pearson *r*. The top heat map shows human vs. automated evaluation correlation across system means, while the bottom heat map shows the correlation across sentences.

MUSS+_matched received the highest scores on average, although they were, on average, not significantly different from the human scores. Note that a small percentage of the human sentences did not have simplifications, which pulled down the average score. As expected, given the automatic complexity measures, those with human-matched target parameters produced simpler sentences than those set to the default settings.

Interrater reliability as measured by quadratic weighted kappa was substantial for all dimensions, with values of $\varkappa$=.65 for Adequacy and $\varkappa$=.74 for Fluency and Simplicity.

4.3 Correlation with human judgment

We computed the Pearson correlation coefficient *r* between human scores and automated metrics for both model mean scores and sentence-level data.

The top heat map in Figure 4 shows mean human scores per system compared with automated metric means on the same subset of data. BERTScore shows the strongest overall correlation across the three human quality measures, with Simplicity corresponding most closely with Precision (*r*=.95) and Adequacy with Recall (*r*=.97), and Fluency with the overall F1 score.

SARI also showed strong positive correlations across categories (*r*=.75-.89) despite the use of only one human reference for comparison. BLEU had very weak correlation (*r*<.10). Unsurprisingly, the compression ratio had a strong negative correlation with Simplicity ratings (higher Simplicity score means simpler and most likely shorter).

The bottom heat map in Figure 4 shows correlations at the sentence level. These correlations show a similar pattern but are overall weaker, suggesting that either there was inconsistency across human ratings or that the automated metrics may be best used to describe entire corpora rather than individual items.

4.4 Comparison with previous work

Martin et al. (2020) reported SARI scores for MUSS+ ranging from 41 to 44. Our results for MUSS+_matched are consistent with these results on our corpus even without multiple human references for comparison. This suggests that MUSS+ performance our new corpus, when matched to characteristics of the human corpus, is comparable to results with Wikilarge and Newsela.

5 Conclusion

In this paper, we present a small parallel corpus for evaluating ATS systems on classic literature and share results on state-of-the-art systems. Our results suggest that the MUSS model trained on both unlabeled data and a supplemental parallel corpus produces sentence simplifications that are similar to humans. The MUSS system also demonstrates the ability to control simplification by changing the target compression ratio while still maintaining text quality and meaning. BERTScore was found to have the strongest correlation with human ratings; in particular the BERTScore Precision strongly correlated with human judgments of simplicity, a key dimension for ATS.

The current work is preliminary and limited due to the size of the corpus and number of references it includes. More robust findings can be found with a larger data set, greater number of human references, and a more singular target use/user for the corpus.

A major issue that became apparent while reviewing literature and producing the corpus is the malleable definition of "simplified" for the purposes of generating and evaluating a corpus. The degree and process of simplification varies across data sets (Brunato et al., 2022), and it is difficult to know how valuable the simplifications will be for their benficiary populations without extrinsic evaluation (Štajner, 2021).

The same questions can be raised for the practical use of controllable systems. While it is clear that MUSS can do some level of adjustment, it is yet to be determined how they can be applied effectively in practice. Similarly, it will be challenging to perform automated evaluations while tweaking these parameters if the intended target ratios greatly differ from the references in the test set. This is another opportunity for more robust human evaluation, both intrinsic and extrinsic.

References

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

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

Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. 2022. Linguistically-based comparison of different approaches to building corpora for text simplification: A case study on italian. *Frontiers in Psychology*, 13.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *ArXiv*, abs/2006.14799.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Project Gutenberg. www.gutenberg.org.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

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

Emma Laybourn. 2017. English literature ebooks, www.englishliteratureebooks.com.

Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2020. Muss: multilingual unsupervised sentence simplification by mining paraphrases. *arXiv preprint arXiv:2005.00352*.

Louis Martin, Benoît Sagot, Eric de la Clergerie, and Antoine Bordes. 2019. Controllable sentence simplification. *arXiv preprint arXiv:1910.02677*.

Constantin Orăsan, Richard Evans, and Ruslan Mitkov. 2018. Intelligent text processing to help readers with autism. In *Intelligent Natural Language Processing: Trends and Applications*, pages 713–740. Springer.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

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

Luz Rello, Ricardo Baeza-Yates, Laura Dempere-Marco, and Horacio Saggion. 2013. Frequent words improve readability and short words improve understandability for people with dyslexia. In *IFIP Conference on Human-Computer Interaction*, pages 203–219. Springer.

Horacio Saggion, Sanja Štajner, Stefan Bott, Simon Mille, Luz Rello, and Biljana Drndarevic. 2015. Making it simplext: Implementation and evaluation

of a text simplification system for spanish. *ACM Transactions on Accessible Computing (TACCESS)*, 6(4):1–36.

Advaith Siddharthan. 2014. A survey of research on text simplification. *ITL-International Journal of Applied Linguistics*, 165(2):259–298.

Punardeep Sikka and Vijay Mago. 2020. A survey on text simplification. *arXiv preprint arXiv:2008.08612*.

Sanja Štajner. 2021. Automatic text simplification for social good: Progress and challenges. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2637–2652.

Sanja Štajner, Ruslan Mitkov, and Horacio Saggion. 2014. One step closer to automatic evaluation of text simplification systems. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 1–10.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. Bleu is not suitable for the evaluation of text simplification. *arXiv preprint arXiv:1810.05995*.

Chris Van Der Lee, Albert Gatt, Emiel Van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. *arXiv preprint arXiv:1703.10931*.

A Appendix

Dimension	Brief scale definitions
Fluency	0 = Completely garbled 1 = Challenging to understand but not completely garbled 2 = At least one major grammatical error (or several minor errors) 3 = Awkward phrasing or minor grammatical error 4 = Completely grammatical
Simplicity	0 = Same complexity or greater complexity compared to the original 1 = Minor simplification (e.g., a single simplified word) 2 = Multiple minor simplifications 3 = Notably simpler, but clear opportunities for simplification missed 4 = Major simplification (e.g., sentence splitting or other major syntactic change, simpler vocabulary)
Adequacy	0 = Meaning not preserved 1 = Major overall meaning change, but some original information remains 2 = Major information changed, omitted, or added, but retains *most* of the original meaning 3 = Minor changes in meaning 4 = Sufficiently similar to the original (Note: A sentence with a minor loss of inessential details can still receive a 4)

Table A1: Quality dimension guidelines. Brief definitions to improve human quality estimation calibration.

Model	Sentence
Original	His many legs, pitifully thin compared with the size of the rest of him,waved about helplessly as he looked.
Human	His many thin legs moved around helplessly as he looked.
ACCESS	His many legs, pitifully, compared with the size of the rest of him.
MUSS	His many thin legs, smaller than his body size, waved helplessly as he looked.
MUSS+	His many legs were very thin. They waved about helplessly as he looked.
Original	I wish either my father or my mother, or indeed both of them, as they were in duty both equally bound to it, had minded what they were about when they begot me.
Human	I wish either my father or my mother, or indeed both, had minded what they were about when they begot me.
ACCESS	I wish either my father or my mother, or when they were in duty both equally bound to it, had minded what they were about.
MUSS	I wish my father or mother, or even both of them, had been aware of what they were doing when I was born.
MUSS+	I wish my father or my mother, or both of them, had minded what they were doing when they had me. They were both bound by duty.
Original	But as I am a party interested in the latter my opinion may perhaps have an undue bias.
Human	But as I am involved in them, I may be biased.
ACCESS	But as I am a party interested in the latter my opinion may have an undue bias.
MUSS	However, since I am the latter's party, my opinion may have an bias.
MUSS+	Because I am interested in the latter, my opinion may be biased.

Table A2: Example simplifications. Each model was configured with control tokens set to match the reference corpus.