# Improving Text Simplification with Factuality Error Detection

**Yuan Ma, Sandaru Seneviratne, Elena Daskalaki**
The Australian National University
{u6712879, sandaru.seneviratne, eleni.daskalaki}@anu.edu.au

## Abstract

In the past few years, the field of text simplification has been dominated by supervised learning approaches thanks to the appearance of large parallel datasets such as Wikilarge and Newsela. However, these datasets suffer from sentence pairs with factuality errors which compromise the models' performance. In this study we proposed a model-independent factuality error detection mechanism, considering bad simplification and bad alignment, to refine the Wikilarge dataset through reducing the weight of these samples during training. We demonstrated that this approach improved the performance of the state-of-the-art text simplification model TST5 by an FKGL reduction of 0.33 and 0.29 on the TurkCorpus and ASSET testing datasets respectively. Our study illustrates the impact of erroneous samples in TS datasets and highlights the need for automatic methods to improve their quality.

## 1 Introduction

Text simplification (TS) is a Natural Language Processing (NLP) task that considers the reduction of text's complexity towards increasing its readability and understandability while retaining its original meaning. TS can increase the accessibility of information to a wider audience, including youngsters, those with little literacy, people who are not native speakers, the elderly, and people with disabilities (Inui et al., 2003; Petersen and Ostendorf, 2007; De Belder and Moens, 2010; Suominen et al., 2013). Additionally, numerous studies have also demonstrated that TS can support other NLP tasks as a preprocessing step (Chen et al., 2012; Chatterjee and Agarwal, 2022).

The current TS domain (Zhang and Lapata, 2017; Martin et al., 2020; Omelianchuk et al., 2021) is dominated by fine tuning large sequence-to-sequence language models on existing parallel datasets, the main ones being Wikilarge (Zhang and Lapata, 2017) and Newsela (Xu et al., 2015). However, several studies have revealed that these training datasets suffer from factuality errors. (Xu et al., 2015; Devaraj et al., 2022) Factuality errors occur when the samples provided do not accurately or properly represent the task. In the TS context, two main sources of factuality errors are bad alignment, i.e., loss of content preservation, and bad simplification, i.e., the target sentence is not simpler than the source (Xu et al., 2015). The existence of parallel training samples with factuality errors can impact significantly the performance of the TS models.

In this study, we investigated methods to detect parallel samples with factuality errors in the Wikilarge dataset. We explored the impact of decreasing the loss weight of the detected samples during training in the TS task performance. We re-trained the state-of-the-art (SOTA) TS model TST5 (Sheang and Saggion, 2021) using the modified Wikilarge dataset and observed a significant performance improvement when tested on the TurkCorpus and ASSET datasets.

## 2 Related Work

### 2.1 Text simplification

Text simplification is mostly treated as a monolingual translation problem based on existing parallel datasets including Wikilarge and Newsela. While previous models focused on using statistical machine translation (SMT) approaches (Coster and Kauchak, 2011; Wubben et al., 2012; Štajner et al., 2015), current work focuses on using neural machine translation (NMT) approaches (Nisioi et al., 2017; Shen et al., 2017; Zhao et al., 2018; Martin et al., 2020). The Neural Text Simplification (NTS) model proposed by Nisioi et al. (2017) is one of the earliest attempts to apply NMT on TS and showed better performance than other SMT models at that time. After the release of transformers, Zhao et al.

*Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 173 - 178
December 8, 2022 ©2022 Association for Computational Linguistics

| | |
|---|---|
| **Bad Alignment** | Complex: They take up oxygen in the lungs or gills and release it while squeezing through the body 's capillaries .<br>Simple: Red blood cells are very large in number ; in women , there are 4.8 million red blood cells per microliter of blood . |
| **Bad Simplification** | Complex: He travelled to Brittany in 1928 to study stone crosses and publish As Cruces de Pedra na Bretaña .<br>Simple: Two years later he published Cousas , and in 1929 he travelled to Brittany to study its stone crosses and publish As Cruces de Pedra na Bretaña . |
| **Real Simplification** | Complex: In September 1869 , O'Reilly escaped and was rescued by an American ship .<br>Simple: In September 1869 , O'Reilly escaped with help from an American ship . |

Table 1: Examples of bad alignment, bad simplification and real simplification in Wikilarge.

(2018) implemented it in their model DCSS and achieved the SOTA performance, highlighting the promising capability of the transformers framework for TS.

Recently, the addition of control tokens was shown to significantly improve the TS models. Martin et al. (2020) proposed one of the currently benchmark models, named ACCESS. Their model included four tokens to control the amount of compression, paraphrase, lexical, and syntactical complexity separately. Later, Sheang and Saggion (2021) improved this method by adding one more token to control the change of sentence length and fine tuning on the pretrained language model T5 (Raffel et al., 2020), resulting in the TST5 model which has achieved the highest reported SARI score on TurkCorpus dataset until now. These works have shown that adding control tokens can significantly improve the performance of TS models.

### 2.2 Factuality errors

Factuality errors happen when sample pairings do not accurately represent the job. They can be divided into two categories: bad simplification and bad alignment (Xu et al., 2015). Bad simplification is identified when the target sentence does not simplify the source sentence, while, when the contents of the source sentence and the target sentence disagree, this corresponds to bad alignment. The topic of factuality errors was addressed by Xu et al. (2015) where, through manual examination of 200 sentence pairs from the Parallel Wikipedia Simplification corpus, they found that 33% of sentence pairs were not simplified, and 17% of sentence pairs were not aligned. Thus, they suggested that Simple Wikipedia was a poor training resource and advised using the Newsela dataset instead. However, Devaraj et al. (2022) recently performed a manual quantitative analysis on both Newsela and Wikilarge and demonstrated that, although Newsela dataset made more proactive simplification operations, it faced a more serious problem with bad simplification error.

## 3 Factuality error detection

In this study, we implemented a rule-based algorithm to detect factuality errors in the Wikilarge dataset. For the detected samples, the loss of the TS model was subsequently scaled down during training to reduce their impact on the model's learning performance.

To detect bad simplification, we utilized the Flesch–Kincaid grade level (FKGL) metric (Kincaid et al., 1975), which was designed for evaluating text readability and has also been used as an evaluation metric in multiple previous works (Martin et al., 2020; Sheang and Saggion, 2021; Omelianchuk et al., 2021). FKGL was originally calculated at the paragraph level based on the average length of the sentence ($\frac{N_{words}}{N_{sentences}}$) and the number of syllables ($\frac{N_{syllables}}{N_{words}}$). To apply FKGL to the sentence level, instead of calculating the average length, the length of the sentence itself was used (Eq. 1). With the assumption that readability reflected simplicity, any sentence pairs for which the source sentence $y$ had higher FKGL score than its target counterpart $x$ were marked as bad simplification pairs.

$$FKGL = 0.39N_{words} + 11.8\frac{N_{syllables}}{N_{words}} - 15.59 \quad (1)$$

Bad alignment was recognized based on named entity recognition. Named entity refers to a phrase that clearly identifies one item from a set of other items that have similar attribute. We identified locations, name, time, and organization in both target and source sentences. This was performed through a pretrained classifier provided by the NLTK library[1](Bird and Loper, 2009). Here, we assumed that simplification might reduce but should not add entities. According to this, we calculated the cosine similarity between all the entities in source sentences and target sentences. Because each named entity may contain different number of words, we used a contextual embedding model based on transformers to create embeddings for each named entity rather than a word level encoder such as word2vec (Mikolov et al., 2013). Bad alignment was recognized if there existed an entity $e_t$ in the target sentence that did not have a corresponding entity $e_s$ in the source sentence with cosine similarity higher than a predefined threshold $T$.

For the pairs marked with factuality error, their corresponding weights were scaled down during training as shown in Eq. 2 and 3 for the bad simplification and bad alignment respectively. The effect of the factuality error samples suppression was explored by experimenting with different scaling parameters $\alpha_1$ and $\alpha_2$.

$$w_1 = \begin{cases} \alpha_1 & \text{if FKGL(x)<FKGL(y)} \\ 1 & \end{cases} \quad (2)$$

$$w_2 = \begin{cases} 1 & \text{if } \forall_{e_t} \exists_{e_s} cos(e_t, e_s) > T \\ \alpha_2 & \end{cases} \quad (3)$$

The resulting weights of bad simplification and bad alignment were multiplied together, and the outcome was then normalized by the total weight. Thus, sentence pairs that were found to be both unaligned and unsimplified were further suppressed.

$$loss = \frac{\sum CrossEntropy(output, label) w_1 w_2}{\sum w_1 * w_2}. \quad (4)$$

## 4 Experiment

### 4.1 Model

We used the TST5 model to evaluate the efficiency of our approach (Sheang and Saggion, 2021). All the training details were unchanged. The T5-based pretrained model was used as the backbone. Huggingface Transformers library[2] and Pytorch-lighting[3] were used to train the model. NLTK library was used for named entity recognition. Huggingface's sentence encoder all-MiniLM-L6-v2[4] was used to create embeddings for named entities. For comparing cosine similarities, the threshold $T$ was set to 0.6, which was selected after experimenting with different threshold values.

In order to enable controllable simplicity, four control tokens were implemented, including NBChars, LevSim, WordRank, and DepTreeDepth, which were identical to ACCESS (Martin et al., 2020). During testing, the control tokens that produced the highest SARI score in the validation set were used.

We investigated different values for the parameters $\alpha_1$ and $\alpha_2$ to explore the impact of the error samples suppression in the model's performance. Specifically, we assessed the model's performance when bad simplification or bad alignment detection was considered with 50% suppression ($\alpha_1/\alpha_2$ = 0.5), 80% suppression ($\alpha_1/\alpha_2$ = 0.2), 98% suppression ($\alpha_1/\alpha_2$ = 0.02), and 100% suppression ($\alpha_1/\alpha_2$ = 0).

### 4.2 Datasets

We used WikiLarge for training and TurkCorpus and ASSET for validation and testing. The three datasets are described below.

**WikiLarge** (Zhang and Lapata, 2017): Contains $29,6402$ sentence pairs from Simple Wikipedia and normal Wikipedia. It is the largest and the most commonly used TS dataset.

**TurkCorpus** (Xu et al., 2016): Contains $2,000$ sentence pairs for validation and 359 sentence pairs for testing. Each sentence has 8 references manually simplified by different people.

**ASSET** (Alva-Manchego et al., 2020): Contains $2,000$ sentence pairs for validation and 359 sentence pairs for testing with 10 references.

[1] https://www.nltk.org

[2] https://huggingface.co/transformers/model_doc/t5.html

[3] https://pytorchlightning.ai

[4] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

|  | TurkCorpus |  |  | ASSET |  |  |
|---|---|---|---|---|---|---|
|  | SARI↑ | FKGL↓ | BLEU↑ | SARI↑ | FKGL↓ | BLEU↑ |
| TST5(Sheang and Saggion, 2021) | 42.46 | 6.28 | 64.26 | 45.17 | 6.31 | 70.04 |
| + bad simplification detection ($\alpha_1 = 0.2$) | 43.06 | 6.12 | 66.07 | 44.75 | 6.19 | 70.87 |
| + bad simplification detection ($\alpha_1 = 0.02$) | 42.87 | 6.08 | 65.50 | 45.10 | 6.29 | 71.42 |
| + bad alignment detection ($\alpha_2 = 0.2$) | 42.84 | 6.38 | 65.93 | 45.17 | 6.27 | 71.04 |
| + bad alignment detection ($\alpha_2 = 0.02$) | 42.90 | 6.15 | 64.91 | 45.03 | 6.23 | 69.42 |
| + both ($\alpha_1, \alpha_2 = 0.5$) | 42.89 | 6.17 | 64.48 | 44.96 | 6.33 | 70.23 |
| + both ($\alpha_1, \alpha_2 = 0.2$) | 43.03 | 5.95 | 64.97 | 45.51 | 6.01 | 70.03 |
| + both ($\alpha_1, \alpha_2 = 0.02$) | 43.25 | 5.95 | 68.32 | 45.12 | 6.02 | 74.03 |
| + both ($\alpha_1, \alpha_2 = 0$) | 43.25 | 6.19 | 67.74 | 45.23 | 6.28 | 72.55 |

Table 2: Performance of the TST5 model trained on the original and modified versions of the Wikilarge and tested on TurkCorpus and ASSET datasets.

To the best of our knowledge, all three datasets were created ethically and are publicly available. No new text data were collected or created as part of this study.

### 4.3 Evaluation metrics

We evaluated the TST5 model's performance using the SARI, FKGL, and BLEU metrics described below.

**SARI** (Xu et al., 2016): Averages F1 scores for addition, keep, and deletion operations with references.

**FKGL** (Kincaid et al., 1975): Evaluates the readability of a sentence.

**BLEU** (Papineni et al., 2002): Assesses how well one sentence matches multiple references.

As SARI is the most adopted metric for TS we used it as our primary metric while FKGL was used to evaluate the simplicity of our output. Although research has shown that BLEU is not suitable for the TS task (Sulem et al., 2018), we included it in our analysis for comparison with previous works. The Wilcoxon signed-rank test (Wilcoxon, 1992) was used to assess the statistical significance of our results.

### 4.4 Results

Our proposed factuality error detection algorithm identified $68,237$ (23 %) samples with bad simplification and $93,030$ (31 %) samples with bad alignment. In total, 45% of the total samples of Wikilarge were identified as factuality errors. The proposed dataset modification with the suppression of both bad simplification and bad alignment samples by factors of $\alpha_1, \alpha_2 = 0.02$ resulted in the best statistically significant improvement of the SARI and FKGL scores by 0.79 and 0.33 respectively on TurkCorpus and improvement of FKGL by 0.29 on ASSET ($p<0.05$). The SARI score on ASSET showed an inconsistent variation, in most of the cases without statistically significant change.

It should be noted that TST5 reported a higher SARI score in the original study(Sheang and Saggion, 2021), but we were unable to reproduce the same results using the code provided by the authors.

## 5 Discussion

Our factuality detection rate was aligned with the work of Xu et al. (2015)'s experiment on the bad simplification case (23% and 33% respectively), however, it identified a higher number of bad alignment samples (31% in comparison to 17%). This could be due to sensitivity differences between the two approaches.

Our TS results (Table 2) demonstrated that the TST5 model's performance could be enhanced by both bad simplification and bad alignment detection. The combination of both factuality errors detection led to improved results. We observed a significant improvement of SARI on TurkCorpus, but not in ASSET, where the SARI score showed an inconsistent but not statistically significant variation. The reason might be due to the SARI score on ASSET being so close to the reference that it was difficult to improve. These results indicate that that the TST5 model trained on the modified Wikilarge was able to generate simpler sentences compared to the original TST5.

From Table 2, it can also be seen that the model's performance improved as the factuality error sample weights decreased. This indicates that the impact of the erroneous samples in the training performance might be more significant than the reduction of the dataset size.

Our results illustrate that the existence of factuality errors in the training datasets used for TS, can induce a significant impact in the performance of the TS models. This indicates a general need for new reliable datasets exploration. Better error detection methods, including more thorough tuning, and further validation is needed with other TS models and other parallel datasets such as Newsela, which is part of our future work. The trade-off between error detection sensitivity and dataset size reduction is crucial and needs further investigation.

## 6 Conclusion

In this paper, we designed a model-independent factuality error detection mechanism to support TS model training. We demonstrated that our mechanism could significantly improve the performance of the SOTA TS model (TST5) based on recognized TS metrics. Our study raises the need for high quality parallel datasets, as well as automated factuality error detecton methods to improve the performance of TS models.

## 7 Limitations

We focused on the Wikilarge dataset and did not include investigation on the Newsela dataset due to lack of access to it at the time of the study. Additionally, we tested our approach on the SOTA TS model TST5 only. However, more models should be tested to assess the generalization of the proposed method. Due to time and resource limitations, we only analyzed our model based on established TS metrics and did not conduct a human evaluation.

## 8 Acknowledgments

This research was funded in part by and has been delivered in partnership with Our Health in Our Hands, a strategic initiative of the Australian National University (ANU), which aims to transform health care by developing new personalized health technologies and solutions in collaboration with patients, clinicians, and health care providers. We wish to acknowledge the support of the ANU School of Computing for Ms Seneviratne's PhD scholarship. We wish to sincerely thank Prof. Hanna Suominen for her valuable support and comments. We also thank NCI Australia (National Computational Infrastructure) for providing computational resources for the project ij84 and the project yv67.

## References

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020. Asset: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. *arXiv preprint arXiv:2005.00481*.

Klein E. Bird, S. and E. Loper. 2009. *Natural language processing with Python*. O'Reilly.

Niladri Chatterjee and Raksha Agarwal. 2022. Studying the effect of syntactic simplification on text summarization. *IETE Technical Review*, pages 1–12.

Han-Bin Chen, Hen-Hsen Huang, Hsin-Hsi Chen, and Ching-Ting Tan. 2012. A simplification-translation-restoration framework for cross-domain smt applications. In *Proceedings of COLING 2012*, pages 545–560.

Will Coster and David Kauchak. 2011. Learning to simplify sentences using Wikipedia. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 1–9, Portland, Oregon. Association for Computational Linguistics.

Jan De Belder and Marie-Francine Moens. 2010. Text simplification for children. In *Proceedings of the SIGIR workshop on accessible search systems*, pages 19–26. ACM; New York.

Ashwin Devaraj, William Sheffield, Byron Wallace, and Junyi Jessy Li. 2022. Evaluating factuality in text simplification. pages 7331–7345.

Kentaro Inui, Atsushi Fujita, Tetsuro Takahashi, Ryu Iida, and Tomoya Iwakura. 2003. Text simplification for reading assistance: a project note. In *Proceedings of the Second International Workshop on Paraphrasing*, pages 9–16.

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.

Louis Martin, Éric de la Clergerie, Benoît Sagot, and Antoine Bordes. 2020. Controllable sentence simplification. pages 4689–4698.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

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

Kostiantyn Omelianchuk, Vipul Raheja, and Oleksandr Skurzhanskyi. 2021. Text Simplification by Tagging. pages 11–25.

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

Sarah E Petersen and Mari Ostendorf. 2007. Text simplification for language learners: a corpus analysis. In *Workshop on speech and language technology in education*. Citeseer.

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.

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.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. *Advances in neural information processing systems*, 30.

Sanja Štajner, Iacer Calixto, and Horacio Saggion. 2015. Automatic text simplification for Spanish: Comparative evaluation of various simplification strategies. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 618–626, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

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

Hanna Suominen, Sanna Salanterä, Sumithra Velupillai, Wendy W Chapman, Guergana Savova, Noemie Elhadad, Sameer Pradhan, Brett R South, Danielle L Mowery, Gareth JF Jones, et al. 2013. Overview of the share/clef ehealth evaluation lab 2013. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 212–231. Springer.

Frank Wilcoxon. 1992. Individual comparisons by ranking methods. In *Breakthroughs in statistics*, pages 196–202. Springer.

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*, 3:283–297.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing Statistical Machine Translation for Text Simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594, Copenhagen, Denmark. Association for Computational Linguistics.

Sanqiang Zhao, Rui Meng, Daqing He, Andi Saptono, and Bambang Parmanto. 2018. Integrating transformer and paraphrase rules for sentence simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3164–3173, Brussels, Belgium. Association for Computational Linguistics.