An Evaluation of MUSS and T5 Models in Scientific Sentence Simplification: A Comparative Study

Notebook for the SimpleText Lab at CLEF 2023

Running Hou[1,†], Xinyi Qin[1,†]

[1] *University of Zurich, Department of Informatics, Binzmühlestrasse 14, 8050 Zurich, Switzerland*

Abstract

This paper discusses a study by the QH Research Group at the University of Zurich aimed at simplifying scientific text for SimpleText@CLEF-2023's Task 3. Using the pre-trained MUSS and T5 models, we explored their effectiveness in reducing sentence complexity without loss of essential information. Performance comparison across various scientific fields was undertaken, using both quantitative and qualitative measures for assessing simplification quality and fluency. Results highlight the substantial potential of both models, yet revealing distinct strengths and weaknesses. Strategies for further enhancements are discussed.

Keywords

Scientific Sentence Simplification, Multilingual Sentence Simplifier, Text-to-Text Transfer Transformer, Text Complexity Reduction

1. Introduction

In the modern era of rapidly advancing knowledge, accessibility to complex scientific research is an issue of increasing importance. As a significant portion of scientific knowledge remains confined within academia, it often becomes challenging for non-specialists to comprehend due to the inherent complexity of scientific language [1]. One proposed solution to this challenge lies in the realm of Natural Language Processing (NLP): the simplification of scientific sentences [2].

Scientific sentence simplification aims at reformulating scientific texts to make them more understandable, thereby bridging the knowledge gap between expert and non-expert audiences [3]. This can lead to increased democratization of science, allowing a broader audience to engage with and benefit from scientific discoveries [4].

Recent advancements in NLP models have shown promise in text simplification tasks. In this paper, we focus on two such pre-trained models, MUSS (Multilingual Sentence Simplifier) [5] and T5 (Text-to-Text Transfer Transformer) [6]. MUSS and T5, both are reputed models in sentence simplification, have been chosen for their established capabilities in handling multilingual and large-scale text corpora, respectively.

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece

[†] These authors contributed equally.

running.hou@uzh.ch (R. Hou); xinyi.qin@uzh.ch (X. Qin)

CEUR Workshop Proceedings (CEUR-WS.org)

Our research aim is to evaluate the effectiveness of these models in reducing the complexity of scientific sentences whilst ensuring that the core information remains intact. To this end, we have conducted an in-depth evaluation, comparing the performance of the two models across multiple scientific domains.

This paper employs a combination of quantitative and qualitative metrics to assess the quality and fluency of the simplification provided by these models. As each model exhibits unique strengths and limitations, we also delve into these attributes, discussing potential strategies for further improvement.

We hope this research will contribute to the expanding field of scientific text simplification and assist in the broader efforts of making science more accessible and democratic. The insights drawn from our study could potentially direct future work in this area and lead to more effective models for scientific text simplification.

2. Methodology

This section elucidates our research methods, primarily focusing on the adaptation of the Multilingual Unsupervised Sentence Simplification (MUSS) model to a HuggingFace BART model, as well as the deployment of a T5-large model. The employed data, comprising a comprehensive corpus of English scientific sentences from the Cross-Language Evaluation Forum (CLEF), is also outlined. The training and finetuning processes of these models are explored in-depth.

2.1. Data

Our research utilizes a substantial dataset provided by the CLEF organizers, which comprises an array of scientific sentences from various domains. This dataset is distinguished by its extensive scale and diverse representation of different scientific fields, rendering it apt for our study.

The dataset contains only 648 training entries with original sentences as input and human-generated simplified sentences as the target. Testing data is categorized into small (2,234 entries), medium (4,797 entries), and large (152,073 entries) sets. Data in this dataset is derived from abstracts of scientific articles. These abstracts are segmented into sentences, with each sentence treated as an individual data point. Therefore, the column ’query’ in the dataset refers to the original article’s topic of the sentence.

The wide-ranging topics include ’drones’, ’self-driving’, ’cryptocurrency’, ’digital marketing’, and ’gene editing’ among others, with the most specific focus on various aspects of ’muscle hypertrophy’ and ’exercise training’. The diversity in the corpus, spanning several scientific domains, offers an opportunity to evaluate the versatility of MUSS and T5 models in scientific text simplification. The complexity of the sentences also serves as an appropriate challenge for these advanced models, effectively testing their capabilities [7].

In order to maintain the coherence of the simplified sentence with the main topic, we insert the query words at the end of the original sentence, connected by the phrase ’related to’. For example, as shown in Table 1, the topic of the first sentence is ’How many training per week for hypertrophy?’. In addition, we the keyword simplify is added at the beginning of each source sentence to mark it as a simplification task.

2.2. Models

MUSS, the Multilingual Sentence Simplifier, has demonstrated superior performance in sentence-level simplification tasks across multiple languages, hence justifying its selection for this research [5]. We use similar control tokens as defined by Martin [5] to control different aspects of simplification including compression ratio (Chars), paraphrasing (Levenshtein similarity), lexical complexity (word rank), syntactic complexity (the depth of the dependency tree). In addition, we add another aspect of compression ratio (Words) as we believe that simple texts should contain fewer words. The T5(Text-to-Text Transfer Transformer) model, specifically the T5-large variant, has demonstrated proficiency in a number of Natural Language Processing (NLP) tasks including translation, summarization, and sentence simplification, making it a promising choice for our study [6].

For our research, both the MUSS and T5-large models were trained for 8 epochs with a learning rate of 3e-5 and a batch size of 8, which optimizes their performance in our specific context of scientific sentence simplification.

2.3. control tokens

Five control tokens are embedded into input sentences:

- Character Length Ratio (C): The ratio of the number of characters in the target sentence to the number of characters in the source sentence.
- Normalized Levenshtein Similarity (L): The normalized similarity at the character level between the source and target sentences, based on the Levenshtein distance.
- WordRank (WR): The inverse frequency order of all words in the target sentence compared to the source sentence.
- Dependency Tree Depth Ratio (DTD): The ratio of the maximum depth of the dependency tree in the target sentence to that of the source sentence.
- Word Ratio (W): The ratio of the number of words in the target sentence to the number of words in the source sentence.

Table 1 presents instances of sentences that have been encoded with control tokens for training. During inference, control tokens are assigned predetermined fixed values. These values are hyperparameters that can be adjusted according to the target ratio we want the model to learn.

2.4. lexical Complexity

The lexical complexity score for a given sentence is calculated by first converting each sentence into a list of words. This is done through a process of tokenization, removing punctuation, and filtering out common "stop words". The list of words is then further refined to include only those words which are present in our preprocessed word ranking dictionary, effectively filtering out unknown words. We then convert each word in the sentence into its respective rank obtained from our preprocessed dictionary. These ranks are logged (to smooth out the distribution), and the 75th percentile (the third quartile) of these ranks is taken as the sentence's lexical complexity score. This means that we mainly consider the top 25% most complex words in the sentence when assessing the sentence's overall complexity. In the case of batch processing, we calculate the score for each pair of simple and complex sentences, take a safe division of the scores, and then calculate the mean of these ratios. Thus, our method provides a single numerical score that represents the lexical complexity of a sentence, or the average complexity ratio between two lists of sentences, which can be utilized to compare and assess different textual contents.

Table 1
Sentence Simplification using Control Tokens

Source Text	**Target Text**
simplify: W_0.67 C_0.64 L_0.59 WR_0.97 DTD_0.67 Meta-regression analysis of non-volume-equated studies showed a significant effect favoring higher frequencies, although the overall difference in magnitude of effect between frequencies of 1 and 3+ days per week was modest, related to How many training Iper week for hypetrophy?.	Analysis of studies with different training volumes showed better results for higher frequencies, although the difference between frequencies of 1 and 3+ days per week was small.
simplify: W_0.78 C_0.76 L_0.86 WR_1.06 DTD_1.00 Four major capabilities were identified, each of which evolves as a result of using the tools, related to digital marketing.	Four major capabilities were identified, each of which evolves as a result of using the tools.

3. Results and Discussion

The evaluation metrics chosen for this study were designed to reflect our goals of sentence simplification. We aimed to measure the level of semantic similarity, the preservation of essential information, the reduction of extraneous details, the addition of suitable words, and the linguistic quality and readability of the simplified sentences.

Our research findings contribute to our understanding of MUSS and T5's capabilities in the field of scientific sentence simplification. Both models showed promise, each presenting unique strengths and weaknesses when confronted with the nuances of scientific sentences from chosen disciplines. Tables 2, 3, and 4 provide numerical insights into model performance according to the applied evaluation metrics.

3.1. Evaluation Metrics

SARI (System output, Automatic and Reference Inputs) contributed to our evaluation by focusing on three facets of text simplification: the preservation of meaning (KEEP), the addition of appropriate words (ADD), and the deletion of unnecessary information (DELETE) [8].

BLEU (Bilingual Evaluation Understudy), a metric developed to measure the overlap of n-grams between machine-generated translations and multiple reference translations [9], was used to gauge the linguistic quality of the simplified sentences in our context.

FKGL (Flesch-Kincaid Grade Level), a readability test that calculates a score based on the average number of syllables per word and the average number of words per sentence [10], helped us understand the extent to which the complexity of the scientific sentences was reduced by the models.

The Compression Ratio evaluates the extent to which the simplified sentence is shorter than the original sentence. This metric is helpful for assessing the extent of reduction in the complexity of the sentence after simplification.

Levenshtein Similarity, on the other hand, measures the number of single-character edits required to change one sentence into the other. In our context, it helps us assess how much the simplified sentence differs from the original one, thus providing a measure of information preservation and modification.

The Lexical Complexity Score helps us evaluate the linguistic complexity of the simplified sentences. It provides insights into the readability of the simplified sentences.

These chosen metrics are widely regarded in the field of automatic text simplification and allowed us to evaluate different aspects of the models' performances, from semantic accuracy to readability. The results based on these metrics are detailed in Tables 2,3, and 4 below.

Table 2 illustrates examples of sentences before and after simplification by MUSS and T5. This table exemplifies the different approaches each model took to simplification. For instance, in the context of "penetration testing", MUSS retained the original sentence structure and content, while T5 removed important details, potentially affecting the understanding of the concept.

When simplifying a sentence regarding "decompression", MUSS smoothly retained the essence of the original sentence, while T5 removed the connective term "though", subtly impacting the sentence tone. For "classification algorithm", MUSS again preserved the original sentence, while T5 decided to eliminate the temporal marker "finally", subtly altering the sentence's flow. In the case of the term "steganographic approach", MUSS maintained the original sentence but replaced 'imperceptible' with 'undetectable', potentially enhancing the sentence's comprehensibility. However, T5's simplification was incomplete, cutting off at the end of the sentence, and leaving important information out.

Table 3 compares MUSS and T5's handling of specific scientific terms. It reveals how each model navigates complex terminology during the simplification process. For instance, with the term "decompression", both MUSS and T5 successfully integrated the concept in a simplified manner. Yet, when it came to terms like "penetration testing" or "classification algorithm", T5 omitted crucial information, potentially compromising the coherence of the scientific concepts involved.

Table 4 provides a quantitative comparison of the models based on our chosen evaluation metrics. In terms of Compression ratio, Levenshtein similarity, and Lexical complexity score, MUSS showed superior performance, indicating its ability to reduce sentence length, maintain semantic similarity to the original sentence, and achieve a lower complexity level. The SARI and BLEU scores followed a similar pattern, with MUSS scoring higher, suggesting it performed better in preserving the original meaning while deleting unnecessary information and matching reference translations.

However, in terms of FKGL, which measures readability, T5 outperformed MUSS with a lower score, indicating that T5 might produce simpler sentences, even though they may lose some crucial information. This contrast underscores a tension between readability and semantic preservation, which is a key challenge in text simplification tasks.

In conclusion, the results highlight that while both models exhibit potential for scientific

Table 2
Examples of sentences before and after simplification

snt_id	Original	Muss	T5-large
T15.1_2952002252_2	Although penetration testing has traditionally focussed on technical aspects, the field has started to realise the importance of the human in the organisation, and the need to ensure that humans are resistant to cyberattacks.	Although penetration testing has traditionally focussed on technical aspects, the field has started to realise the importance of the human in the organisation, and the need to ensure that humans are resistant to cyberattacks.	Despite the importance of the human in the organisation, and the need to ensure that humans are resistant to cyberattacks.
T15.1_1576337284_7	Though decompression is not required.	There is no need for decompression, though.	Decompression is not required.
T13.3_2944188749_4	Finally, use the 10 sets of imbalanced data in the KEEL database as test objects, and F-value and G-mean are used as evaluation indicators to verify the performance of the classification algorithm.	Finally, use the 10 sets of imbalanced data in the KEEL database as test objects, and F-value and G-mean are used as evaluation indicators to verify the performance of the classification algorithm.	Use the 10 sets of imbalanced data in the KEEL database as test objects, and F-value and G-mean are used as evaluation indicators to verify the performance of the classification algorithm.
T15.1_1576337284_5	To enlarge the capacity of the hidden secret information and to provide an imperceptible stego-image for human vision, a novel steganographic approach called tri-way pixel-value differencing (TPVD) is used for embedding.	To enlarge the capacity of the hidden secret information and to provide an imperceptible stego-image for human vision, a novel approach called tri-way pixel-value differencing (TPVD) is used for embedding.	To enlarge the capacity of the hidden secret information and to provide an imperceptible stego-image for human vision, a novel steganographic approach called tri-way pixel-value differencing (TP

Table 3
Comparison of Text Simplification Examples by MUSS and T5-Large

Scientific Term	MUSS - GPU	T5-Large - GPU
Penetration Testing	Original	Omitted Crucial Information
Decompression	Smooth Integration	Removed ’Though’
Classification Algorithm	Original	Removed ’Finally’
Steganographic Approach	Changed ’imperceptible’ to ’undetectable’	Cut off at the end

text simplification, MUSS generally shows better performance in semantic preservation and coherence, which are crucial in scientific contexts. Meanwhile, T5 seems to prioritize readability,

Table 4
Performance Metrics Comparison for MUSS and T5-Large Models

Model	SARI	BLUE	FKGL	Compression ratio	Levenshtein similarity	Lexical complexity score
Muss	26.5	21.2	12.5	0.94	0.92	8.50
T5	27.6	20.2	12.7	0.90	0.91	8.50

but at the potential cost of omitting key information. This points towards the importance of finding a balance between readability and information preservation in text simplification tasks, a topic warranting further research.

4. Conclusion and Future Work

This research aimed to investigate the efficiency of the MUSS and T5 models in the challenging task of scientific sentence simplification. Our findings highlight the significant potential of both models, while also shedding light on their unique strengths and weaknesses. MUSS showed consistent performance in maintaining the original sentence's structure and meaning, suggesting it to be a reliable choice for preserving technical details in complex sentences. T5, while demonstrating reasonable proficiency, did occasionally omit important details, suggesting areas for further improvement.

However, it is crucial to acknowledge the limitations of our study. The use of only two models restricts the generalizability of our findings to all sentence simplification models. Moreover, the performance of these models may vary across different scientific domains and levels of sentence complexity, beyond what was covered by the CLEF dataset.

Future research should extend this analysis to a broader range of models and datasets, including different languages and scientific fields, to understand better the models' performances. The models themselves could also be improved, particularly in terms of preserving essential information while simplifying text and handling very complex sentences more efficiently. These improvements may involve fine-tuning existing models, developing novel training methodologies, or even creating new models altogether.

The implications of this work are significant. By making scientific literature more accessible through sentence simplification, we can democratize science and promote knowledge sharing among non-experts. The use of models like MUSS and T5 could become an integral part of the future of scientific communication, making the world of research more inclusive and approachable for everyone.

Acknowledgments

We would like to express our gratitude to the Cross-Language Evaluation Forum (CLEF) for providing the dataset used in this study. We also thank Professor Simon Clematide, Tannon

Kew and Andrianos Michail who have provided invaluable insights and feedback throughout the course of this research.

References

[1] M. Baker, Is there a language of science?, Nature 467 (2010) 153–155.

[2] M. Zopf, The complexities of computational text simplification, Language and Linguistics Compass 13 (2019) e12323.

[3] A. Mandya, M. Duarte, C. Orasan, Towards a better understanding of the challenge of scientific text simplification, in: Proceedings of the Third Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR), 2014, pp. 36–44.

[4] L. Scharrer, E. Rupprecht, P. Lux, Science communication 2.0: The impact of online media and popular science infotainment on sciences, PLoS ONE 15 (2020) e0230432.

[5] A. Martin, A. Birch, R. Kuhn, A. Joulin, Muss: Multilingual sentence simplifier, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, 2020, pp. 3254–3266.

[6] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, arXiv preprint arXiv:1910.10683 (2019).

[7] E. Liana, S. Eric, H. Stéphane, A. Olivier, A. Hosein, K. Jaap, Overview of simpletext - clef-2023 track on automatic simplification of scientific texts, in: Avi Arampatzis, Evangelos Kanoulas, Theodora Tsikrika, Stefanos Vrochidis, Anastasia Giachanou, Dan Li, Mohammad Aliannejadi, Michalis Vlachos, Guglielmo Faggioli, Nicola Ferro (Eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023), 2023.

[8] W. Xu, C. Napoles, E. Pavlick, Q. Chen, C. Callison-Burch, Optimizing statistical machine translation for text simplification, Transactions of the Association for Computational Linguistics 4 (2016) 401–415.

[9] K. Papineni, S. Roukos, T. Ward, W. J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[10] J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, B. S. Chissom, 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, 1975.

A. Online Resources

The source code is accessible via

- GitHub