Automatic Simplification of Scientific Texts using Pre-trained Language Models: A Comparative Study at CLEF Symposium 2023

Aftab Anjum[1,†], Nikolaus Lieberum[2,†]

[1]*Christian-Albrechts-Universität zu Kiel, Christian-Albrechts-Platz 4, 24118 Kiel, Germany*

Abstract

The complexity of scientific texts often creates a barrier to understanding for non-specialist readers. This barrier inhibits the democratization of knowledge and prevents the wider public from engaging with scientific discourse. We argue that applying artificial intelligence to the tasks of identifying and explaining difficult concepts (Complexity Spotting), and simplifying scientific text, has the potential to democratize access to scientific knowledge. We investigate a range of cutting-edge deep learning models for their efficacy in these tasks. The models are trained and evaluated on a dataset of scientific articles, annotated for complex concepts and their simpler explanations. We present a comparative analysis of the performance of these models, illuminating the strengths and weaknesses of each. Our findings reveal promising avenues for future research and development in the field of automated text simplification, contributing to the broader goal of making scientific knowledge accessible to all.

Keywords

Text Simplification, Natural Language Processing, Difficult Term Extraction, Difficult Term Explanation, Neural Networks

1. Introduction

Scientific texts are known for their complexity, technical jargon, and specialized vocabulary, which can often pose challenges for a wide range of readers. Researchers, scientists, and students alike often struggle to comprehend and digest the content of scientific papers, making it difficult for them to stay up-to-date with the latest advancements in their respective fields. To bridge this gap and promote accessibility of scientific information, there is a growing need for automatic simplification techniques that can transform intricate scientific texts into more comprehensible versions without compromising the integrity and accuracy of the original content. In recent years, significant progress has been made in natural language processing (NLP) and machine learning, enabling the development of various text simplification techniques. One such technique, called SimpleText, focuses specifically on the automatic simplification of scientific texts. SimpleText aims to address the linguistic and structural complexities present in scientific writing while preserving the essential scientific concepts and ensuring the accuracy of the simplified

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

[*] Corresponding author.

[†] These authors contributed equally.

afa@informatik.uni-kiel.de (A. Anjum); stu229910@mail.uni-kiel.de (N. Lieberum)
https://www.zbw.eu/de/forschung/information-profiling-and-retrieval/profil-aftab-anjum (A. Anjum)

content. However, the unique characteristics of scientific texts present additional challenges for automatic simplification. Scientific texts often contain domain-specific terminology, complex sentence structures, and intricate logical reasoning, which require a deep understanding of the underlying concepts. Existing text simplification approaches, designed for general-purpose texts, may not adequately capture the nuanced relationships and meaning in scientific content.

For the Blended Intensive Program (BIP) Artificial Intelligence (AI) for Humanities: from Text Simplification to Automatic Humor Analysis, we explore the application of advanced deep learning models namely AI21, ST5, and BLOOM to address the challenges of text simplification. These models, each with their unique capabilities, are applied to two interconnected tasks: 'Complexity Spotting,' where the objective is to identify and explain difficult concepts in scientific texts, and 'Text Simplification,' where the aim is to convert complex scientific sentences into simpler ones that are easier for a general audience to understand.

The overarching aim of this paper is to investigate the efficacy of these deep learning models in simplifying scientific texts and spotting complex concepts. By doing so, we hope to contribute valuable insights to the field of automated text simplification, ultimately promoting the wider accessibility and understanding of scientific knowledge.

The structure of the paper is as follows: Section 2 provides a concise overview of the related work in the field. Section 3 presents the experiments conducted in this study, including detailed descriptions of the tasks, attributes of the dataset used, a discussion on the utilization of existing models, and a comprehensive analysis of the obtained results. Finally, Section 4 concludes the paper by summarizing the key findings and highlighting the significance of the advancements made. It also addresses the future directions and potential strategies for further improving the performance of the models.

2. Related Work

Automatic text simplification is a research area that focuses on developing computational methods to simplify complex texts and make them more accessible to a wider audience, including individuals with cognitive or linguistic challenges, non-native speakers, or people with low literacy levels. This field combines techniques from natural language processing (NLP), machine learning, and linguistics to analyze and modify the structure, vocabulary, and syntax of texts.

There has been significant research and development in automatic text simplification, aiming to create algorithms and models that can effectively simplify texts while preserving their meaning. Some common approaches i am doing to discuss here.

2.1. Lexical Based automatic text simplification

Lexical-based automatic text simplification is an approach that focuses on simplifying texts by replacing complex words or phrases with simpler alternatives while preserving the overall meaning. This technique leverages lexical resources, such as dictionaries, thesauri, and word frequency lists, to identify and substitute complex terms with simpler equivalents.

The research community has made notable strides in the domain of lexical-based automatic text simplification. For instance, Ciprian-Octavian and Andrei-Ionut Stan [1] introduces SimpLex, a lexical text simplification architecture designed to automatically simplify complex texts.

The architecture focuses on lexical substitutions, where complex words or phrases are replaced with simpler alternatives. SimpLex leverages a combination of linguistic resources, such as WordNet and SimpleWiki, along with machine learning techniques to identify suitable substitutions. The system is evaluated on a large corpus of news articles and achieves significant improvements in readability while preserving essential content. The research demonstrates the effectiveness of a lexical-based approach to text simplification and provides a valuable resource for improving the accessibility of written texts for a wider range of readers.

An other work Proposed Debabrata and Tambe [2] where they used lexical for text simplification approach using WordNet. The authors propose a method that identifies complex words in a given text and replaces them with simpler synonyms from WordNet. The approach involves measuring semantic relatedness between words and selecting the most suitable substitution based on a combination of contextual and lexical cues. Evaluation results demonstrate the effectiveness of the proposed method in improving the readability of complex texts while preserving the core meaning. The research contributes to the field of text simplification by providing a valuable and accessible solution, leveraging the rich lexical information offered by WordNet to automatically simplify complex vocabulary.

In a related context, Jipeng Qiang [3] introduces LSBert, a lexical simplification method based on BERT, a popular language representation model. The authors propose an approach that leverages BERT's contextualized embeddings to generate simplified versions of complex words or phrases. LSBert employs a two-step process: first, it identifies complex words in the input text, and then it generates simpler alternatives by selecting candidate substitutions based on their semantic similarity to the original word. The simplification is performed by fine-tuning BERT on a large corpus of simplified pairs. Evaluation results demonstrate the effectiveness of LSBert in simplifying complex vocabulary while maintaining the overall coherence and meaning of the text. This research contributes to the field by harnessing the power of BERT in lexical simplification tasks and offers a promising solution for enhancing the accessibility and understandability of written content.

Moreover, in 2019 Sanja and Horacio [4] focuses on improving the lexical coverage of text simplification systems specifically designed for the Spanish language. The authors address the challenge of limited lexical resources available for Spanish text simplification by proposing a method that combines rule-based strategies and machine learning techniques. They leverage existing resources, such as WordNet and specialized corpora, to build a comprehensive lexicon specifically tailored for Spanish text simplification. The proposed method effectively identifies complex lexical items and suggests appropriate substitutions. Evaluation results demonstrate that the enhanced lexical coverage significantly improves the performance of Spanish text simplification systems, leading to more accurate and effective simplifications. This research provides a valuable contribution to the field by addressing the lexical challenges specific to the Spanish language and enhancing the accessibility and understandability of Spanish texts for a wider audience.

2.2. Semantic Based automatic text simplification

Semantic-based automatic text simplification is an approach that aims to simplify texts while preserving their underlying meaning and intention. By leveraging semantic analysis, this technique identifies complex linguistic structures and replaces them with simpler alternatives, enhancing the accessibility of the text for a broader readership.

Numerous research studies have delved into the realm of semantic-based automatic text simplification, yielding valuable contributions to the field. For instance, Elior Sulem presents a novel approach [5] to evaluate the effectiveness of text simplification techniques. The authors propose a framework that combines semantic analysis and structural evaluation to assess the quality of simplified texts. The framework considers both the preservation of the original meaning and the improvement in readability. The authors conduct experiments on a large dataset of simplified texts and demonstrate that their approach outperforms existing evaluation methods in capturing both semantic and structural changes. The findings of this study contribute to the development of more accurate and reliable evaluation metrics for text simplification systems, ultimately leading to improved accessibility and comprehension for diverse readers.

In another study by Sanja and goran paper proposes a novel approach [6] to automated text simplification by leveraging event-based semantics. The authors recognize that complex sentence structures pose significant challenges to comprehension, especially for individuals with limited language proficiency. To address this, they introduce a method that focuses on identifying key events and their participants in a sentence. By simplifying sentence structures while preserving the fundamental meaning conveyed by these events, the proposed approach aims to improve the accessibility and understandability of complex texts. Evaluation results demonstrate promising performance, with the method successfully simplifying sentences while maintaining their semantic coherence and preserving critical information. This research provides valuable insights into the use of event-based semantics for text simplification, offering a potentially effective solution to enhance the accessibility of complex texts for various user groups.

Similarly, Shuming and Xu Sun Introduces a method [7] that leverages event-based semantics for automated text simplification. The authors propose an approach that focuses on simplifying complex sentences by representing their semantic structure in terms of events and their participants. By identifying the main event and its semantic roles, the method generates simpler versions of the sentences while maintaining the core meaning. Evaluation results demonstrate that the proposed approach effectively simplifies complex sentences while preserving semantic coherence. This research contributes to the field by providing a novel perspective on text simplification, emphasizing the importance of event-based semantics in simplifying complex texts and making them more accessible to a wider range of readers.

2.3. Transformer models for Automatic text simplification

Transformer models have emerged as powerful tools for automatic text simplification, offering state-of-the-art performance in various natural language processing tasks. In the context of text simplification, transformer models have been widely applied and have shown promising results. Some of the study in this domain i will put here. In 2018, Sanqiang and Rui Meng [8] proposes an approach for sentence simplification that integrates Transformer models with paraphrase rules. The authors acknowledge the challenges of simplifying sentences while maintaining their meaning and grammatical correctness. To address this, they combine the power of Transformer models, known for their ability to learn contextual representations, with manually curated paraphrase rules. The method involves generating multiple simplified versions of a source sentence using the Transformer model and then applying the paraphrase rules to ensure simplicity and coherence. Evaluation results demonstrate that the proposed approach outperforms existing methods in terms of simplicity and grammaticality. This research contributes to the field by presenting a comprehensive approach that combines the strengths of Transformer models and human-created paraphrase rules, offering a promising solution for sentence simplification.

Similarly, Robert-Mihai proposed a study [9] where they explores the application of sequence-to-sequence (Seq2Seq) models for automated text simplification. The authors recognize the importance of enhancing the accessibility of complex texts for individuals with lower reading abilities. To address this, they propose a method where a Seq2Seq model is trained to generate simplified versions of input sentences. The model is trained on a large dataset of sentence pairs, consisting of complex and simplified versions. By learning to map complex sentences to simpler equivalents, the Seq2Seq model offers a promising solution for text simplification. Evaluation results demonstrate that the proposed approach significantly improves readability while maintaining the core meaning of the original text. This research contributes to the field by showcasing the effectiveness of Seq2Seq models for automated text simplification, highlighting their potential to make complex texts more understandable and inclusive.

Moreover, Takumi Maruyama [10] focuses on the challenging problem of text simplification for languages with extremely low resources. The authors propose an approach that leverages pre-trained Transformer-based language models, such as BERT, to overcome the limitations of data scarcity. By fine-tuning these models on small amounts of labeled simplification data, they are able to generate simplified versions of complex sentences. Evaluation results demonstrate the effectiveness of this approach, with the generated simplifications achieving high levels of simplicity and readability. The research demonstrates the potential of pre-trained Transformer models to address low resource scenarios, opening up possibilities for text simplification in languages with limited available data. This work contributes to the field by providing insights into adapting large-scale language models for low resource text simplification, making it a valuable contribution for improving the accessibility of complex texts in low resource language contexts.

3. Experiments

3.1. Task description for CLEF (2023) Simple-Text

The Simple-Text data-set and benchmarks contribute to the research on automatic text simplification by introducing three interconnected tasks.

Task 1: What is in (or out)? Select passages to include in a simplified summary, given a query.

Task 2: What is unclear? Given a passage and a query, rank terms/concepts that are required to be explained for understanding this passage (definitions, context, applications,..).

Task 3: Rewrite this! Given a query, simplify passages from scientific abstracts.

3.2. Dataset

In our research, we leverage the DBLP abstracts corpus as the main source of our data. Specifically, we utilize the Citation Network Dataset, known as DBLP+Citation, which is the 12th version released in 2020. This dataset consists of a vast collection of 4,894,063 scientific articles. By using this comprehensive corpus, we ensure a diverse and extensive range of scholarly publications for our analysis and experimentation. The DBLP+Citation dataset serves as a valuable resource, providing us with the necessary information and content to investigate various research questions and explore the intricacies of scientific articles in our study.

For the task of Complexity Spotting, we rely on an annotated database that includes the following columns: query_id (e.g., G11.1), query_text (e.g., drones), snt_id (e.g., G11.1_2892036907_1), source_snt, term (e.g., autonomous), difficulty, and definition. This database serves as a comprehensive resource for training and evaluating our models, providing diverse examples of complex terms along with their definitions and levels of difficulty.

For the Text Simplification task, we utilize another dataset which includes the columns: query_id, query_text, doc_id, snt_id, source_snt, and simplified_snt. This dataset offers a variety of complex sentences from scientific texts alongside their simplified versions, offering a strong foundation for the evaluation and improvement of our models' simplification capabilities.

3.3. Utilization of Existing Models

In the following, the approach used for the two tasks involved in this research will be illustrated. Detailed procedures for data gathering and subsequent preprocessing are covered, along with the process for extracting features. Moreover, we elucidate the choice and preparation of the machine learning models, coupled with the performance evaluation metrics utilized for its assessment.

SimpleT5 Simple T5 is a model built on top of PyTorch Lightning and Transformers. It allows users to quickly train their T5 models, including T5, mT5, and byT5 models, with only a few lines of code [10]. The T5 models, which can be trained using SimpleT5, are versatile and can be used for a variety of natural language processing (NLP) tasks. These tasks include summarization, question answering (QA), question generation (QG), translation, text generation, and more [11].

AI21 Labs - Jurassic-2 Grande Instruct The J2-Grande-Instruct model is a variation of the Jurassic-2 series developed by AI21. It is an auto-regressive language model based on the Transformer architecture and designed with modifications for improved efficiency. The models diverge from their GPT-3 counterparts in several aspects, including vocabulary size and the depth/width ratio of the neural net. [12] This model is specifically trained to handle instructions-only prompts, also known as "zero-shot" prompts, without the need for examples or "few-shot" prompts. It aims to provide a natural way to interact with large language models and is designed to give users an idea of the optimal output for their task without needing any examples.

BLOOM (BigScience Large Open-science Open-access Multilingual Language Model) The BLOOM model is an autoregressive Large Language Model (LLM) that leverages a decoderonly transformer architecture, derived from Megatron-LM GPT-2. It underwent training on approximately 366 billion tokens between March and July 2022, utilizing 1.6 Terabytes of preprocessed text. This extensive dataset included 350 billion unique tokens, encompassing 46 natural languages and 13 programming languages, enabling BLOOM to grasp a wide range of linguistic and programming contexts [13].

3.4. Results Analysis

In this section, we present a detailed analysis of the results obtained from our experiments. We begin by providing an overview of the experimental setup and methodology employed for the evaluation. Subsequently, we delve into the quantitative analysis of the performance metrics, followed by a qualitative assessment of the generated outputs.

For Bloom, AI21 and T5, the training processes of these models involve large-scale language modeling, leveraging vast amounts of text data. Bloom utilizes a combination of unsupervised and supervised training techniques, incorporating linguistic knowledge and fine-tuning on specific downstream tasks. AI21 adopts a similar approach, employing a Transformer-based architecture and training on a diverse dataset. T5, on the other hand, employs a unified framework that incorporates both supervised and unsupervised learning, enabling it to perform multiple tasks. These pre-trained models can be utilized by fine-tuning them on specific downstream tasks, such as text classification, summarization, or question-answering. By adapting the pre-trained models to target tasks, researchers and practitioners can benefit from their powerful language understanding capabilities and achieve improved performance in a range of NLP applications.

This study we did not optimised the hyper parameters of the above utilized models for simple tasks. we used the default parameters for Task 2.2 and Task 2.1. we focused on utilizing pre-trained language models without performing fine-tuning. The models, including Bloom, AI21, and T5, were accessed through their respective APIs to obtain results for the tasks at hand. Specifically, Table 1 and Table 2 present the performance of the T5 model for Task 2.2. Similarly, for Task 2.1, no fine-tuning was conducted, and the pre-trained models were utilized as is. Upon analyzing the provided tables (1, 2, 3, 4), it becomes evident that the achieved performance is not particularly high. This can be attributed to the fact that the models used are not specifically tailored to the domain of the study. However, it is important to note that by undertaking the process of fine-tuning the models with our specific training data, the results are anticipated to exhibit significant improvements. Fine-tuning the models to align with the domain-specific requirements of the study would enhance their performance and generate more favorable outcomes.

Table 1
Accuracy score on Test Data set of Task 2 (2.2).

Model	total	BLEU	FKGL	SARI
ST5	245	0.21	12.77	27.19

Table 2
Accuracy score on Test Data set of Task 2 (2.2).

Model	total	BLEU	FKGL	SARI
ST5	648	0.60	12.30	65.00

Table 3
Accuracy score on Test Data set of Task 2 (2.1).

Model	total	BLEU	ROUGE_precision	ROUGE_recall	ROUGE_fmeasure	semantic_match
AI21	18	0.07	0.31	0.39	0.32	0.79
BLOOM	9	0.13	0.29	0.28	0.28	0.53
ST5	197	0.03	0.27	0.21	0.22	0.61

Table 4
Accuracy score on Test Data set of Task 2 (2.1).

Model	Accuracy
Ai21	0.43
BLOOM	0.46
ST5	0.80

4. Conclusion

The comprehension of scientific texts can be challenging for non-specialist readers, acting as a barrier that restricts access to scientific knowledge and inhibits public engagement in scientific discussions. To address this issue and foster the democratization of scientific information, we propose leveraging artificial intelligence techniques for identifying and explaining complex concepts (referred to as Complexity Spotting) and simplifying scientific texts. This study investigates state-of-the-art deep learning models for their effectiveness in these tasks. We train and evaluate the models using a data-set of scientific articles that have been annotated to identify complex concepts and their corresponding simpler explanations. Through a comparative analysis, we provide insights into the strengths and weaknesses of each model's performance. Our findings highlight promising opportunities for future research and development in automated text simplification, which contributes to the overarching objective of making scientific knowledge more accessible to a wider audience.

In this study, we did not perform fine-tuning on the large language models such as T5, Bloom, and AI21. Despite this, the performance of the models remained reasonable. For our future investigations, we plan to conduct thorough data exploration and analysis as a preliminary step. Subsequently, we intend to fine-tune these models using the provided dataset. Given that these models have already been trained on extensive data, fine-tuning can be achieved with a smaller amount of additional data. Additionally, these models exhibit some capability in handling data imbalance, albeit to a certain extent, if the dataset is not heavily skewed. By pursuing these steps, we aim to further enhance the performance and applicability of the models in the context of scientific text simplification.

References

[1] C.-O. Truică, A.-I. Stan, E.-S. Apostol, Simplex: a lexical text simplification architecture, Neural Computing and Applications 35 (2023) 6265–6280.

[2] D. Swain, M. Tambe, P. Ballal, V. Dolase, K. Agrawal, Y. Rajmane, Lexical text simplification using wordnet, in: Advances in Computing and Data Sciences: Third International Conference, ICACDS 2019, Ghaziabad, India, April 12–13, 2019, Revised Selected Papers, Part II 3, Springer, 2019, pp. 114–122.

[3] J. Qiang, Y. Li, Y. Zhu, Y. Yuan, Y. Shi, X. Wu, Lsbert: Lexical simplification based on bert, IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021) 3064–3076.

[4] S. Štajner, H. Saggion, S. P. Ponzetto, Improving lexical coverage of text simplification systems for spanish, Expert Systems with Applications 118 (2019) 80–91.

[5] E. Sulem, O. Abend, A. Rappoport, Semantic structural evaluation for text simplification, arXiv preprint arXiv:1810.05022 (2018).

[6] S. Štajner, G. Glavaš, Leveraging event-based semantics for automated text simplification, Expert systems with applications 82 (2017) 383–395.

[7] S. Ma, X. Sun, A semantic relevance based neural network for text summarization and text simplification, arXiv preprint arXiv:1710.02318 (2017).

[8] S. Zhao, R. Meng, D. He, S. Andi, P. Bambang, Integrating transformer and paraphrase rules for sentence simplification, arXiv preprint arXiv:1810.11193 (2018).

[9] R.-M. Botarleanu, M. Dascalu, S. A. Crossley, D. S. McNamara, Sequence-to-sequence models for automated text simplification, in: Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part II 21, Springer, 2020, pp. 31–36.

[10] T. Maruyama, K. Yamamoto, Extremely low-resource text simplification with pre-trained transformer language model, International Journal of Asian Language Processing 30 (2020) 2050001.

[11] S. Roy, simplet5, https://pypi.org/project/simplet5/, 2022. Accessed: 2023-06-05.

[12] O. Lieber, O. Sharir, B. Lenz, Y. Shoham, Jurassic-1: Technical Details and Evaluation, Technical Report, AI21 Labs, 2023. URL: https://uploads-ssl.webflow.com/60fd4503684b466578c0d307/61138924626a6981ee09caf6_jurassic_tech_paper.pdf.

[13] B. Workshop, Bloom: A 176b-parameter open-access multilingual language model, 2023. arXiv:2211.05100.