Is Using an AI to Simplify a Scientific Text Really Worth It?

Léa Talec-Bernard[a]

[a] *University of Western Brittany (UBO), 20 Duquesne Street, Brest, 29490, France*

Abstract

Nowadays, we encounter Artificial Intelligence (AI) everywhere and although it can help us save a lot of time when performing certain tasks such as translating or correcting a text, it can also lead to a loss in quality. When it comes to automatically simplifying scientific works, staying true to the meaning of the text has to stay the most important aspect. Unfortunately, this aspect happens to be one of the hardest to control.

We participated in the tasks 2 and 3 of the SimpleText track. They were carried in order to analyse the quality of automatic simplification. Both of the tasks were done using the simplification model T5 with the use of Pyhton. SimpleText task 2 consisted in automatically identifying complicated terms and ranking them according to their level of difficulty. SimpleText task 3 consisted in simplifying sentences from a scientific text. The final results were far from satisfying. Mots of the sentences were cut after a few words, other stayed unchanged.

Keywords

artificial intelligence; simplification; scientific texts; T5 model

1. Introduction

Nowadays, we are used to encountering AI (Artificial Intelligence) everywhere. Whether it be cars, smartphones or even social media, the latest technologies seem to heavily rely on the use of AI to grant more autonomy to the user and to help them save precious time. Some examples of AI we rarely think about, although known to all, is those created to alter textual documents such as automatic translators, automatic spell checkers or even automatic simplifiers.

According to the computer scientist and writer Horacio Saggion, “automatic text simplification is the process of transforming a text into another text which, ideally conveying the same message, will be easier to read and understand by a broader audience.” [1]. That said, one can easily understand the benefits of simplifying a scientific text.

Most of the time, scientific texts are written for other scientists to read. Therefore, they usually use a very technical language and are also extremely detailed which makes them difficult to read and understand for most people. Simplifying scientific texts could help educate the many people who don’t usually have access to this type of information, it could also help people with reading or comprehension difficulties. Automating that simplification could allow anyone to have access to any scientific text. Simplifying scientific texts does not only serve people but also machines. For example, some texts can be simplified to be better analyzed by machines.

This paper aims to answer the following question: Is using an AI to simplify a scientific text really worth it?. Can it give its user a satisfying result and help them save time and efforts? In order to do so, the transformer model T5 [2] created by Google was put to the test.

CLEF 2022 – Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy
EMAIL: Lea.Talec-Bernard@etudiant.univ-brest.fr


2. Related Work

Many researchers have worked on the topic of the automatic simplification of scientific texts. Rémi Cardon, for example, produced several papers on this topic. One of them, "Lexical approach to automatic simplification of medical texts" ("Approche lexicale de la simplification automatique de textes médicaux".[3]), focuses on the lexical simplification, a simplification which aims at replacing complex words or groups of words by simpler equivalents carrying the same meaning. Remi Cardon also created, alongside Natalia Grabar, a corpus designed to improve the automatic simplification of medical documents in French [4].

As for the English language, the paper "Design, development and validation of a system for automatic help to medical text understanding" [5] illustrates the creation of a simplification tool specifically designed for medical documents.

3. Task

Two tasks were performed in order to analyze the use of the T5 model. More precisely, we participated in the tasks 2 and 3 of the SimpleText track. Task 2 of the SimpleText [6] lab is organized as a part of the CLEF-2022 conference and named "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,..).". The goal of this task is to choose which words in a sentence are susceptible to be poorly understood by their readers. SimpleText task 3 is "Rewrite this! Given a query, simplify passages from scientific abstracts.". As its name suggests, this task aims at simplifying sentences from scientific texts.

The data for the tasks 2 and 3 are divided into two sets. The first one, the "train data" is composed of 453 abstracts about computer science and about medicine. These abstracts come from the Citation Network Dataset: DBLP+Citation, ACM Citation network (12th version) [7] as well as Google Scholar and PuMed articles on muscle hypertrophy and health. Difficult terms were then extracted from these abstracts and attributed a score between 1 and 3 and between 1 and 5 according to their level of complexity (1 being the simplest terms and the heist number available the most complicated terms). The train data is used to train the T5 model in order to obtain results close to those illustrated in this data. The second set, the test data, is composed of 116,763 sentences from the DPLB dataset.

4. Methodology description

The Google T5 Model [2] was executed with the use of Python to perform these tasks with the help of the Simple T5 library[1]. This model is one of the most performant models existing today. It allows its users to perform many different NLP (Natural Language Processing) tasks which include translating, summarizing or simplifying documents amongst others. Its performance comes from the fact that it is pre-trained with a very diverse corpus called C4 (Colossal Clean Crawled Corpus) which gathers data from the Common Crawl, an open corpus created from many documents found on the net especially for projects requiring a lot of data. The data from the C4 corpus has been filtered in order to obtain the best result possible. In addition to this data, it is possible to train the T5 model with our own data similar to the document we want to modify. The user also has the possibility to change some of its parameters, allowing a more personalized result. Some of these parameters include the number of times the model will train on the training data or the creativity of the model.

The data was split into sentences in order to better visualize the results achieved through the T5 model. The results were then evaluated on whether or not they were satisfying.

[1] https://github.com/Shivanandroy/simpleT5

5. Results

The results for this simplification were not satisfying. The vast majority of the target sentences were cut off in the middle. Some other complex sentences stayed unchanged after the use of the T5 model,. In other cases, sentences were slightly changed but stayed complex, in the following example part of the sentence is erased. This part was not of major importance and therefore allows the viewer to focus on the essential meaning of the sentence. However, many words in the target sentence are too complex to be considered part of a simplified version of the text: the words “implementing” or “embedded” for example, could be replaced with simplified synonyms. The “Privacy-by-Design approach” mentioned in this example, could also be described.

Table 1
Example: under-simplification

Source text	Target text
In an attempt to bridge this gap, this paper uncovers hidden issues of the Privacy-by-Design approach as a means to derive privacy requirements for implementing information systems with privacy embedded by design.	This paper uncovers hidden issues of the Privacy-by-Design approach as a means to derive privacy requirements for implementing information systems with privacy embedded by design.

In other examples, sentences loose part of their meaning after being modified by the T5 model. In the following example, the first few words “To this end” are replace by “To address this end”, which not only adds complexity to the sentence but also removes part of its meaning.

Table 2
Example: loss of meaning

Source text	Target text
To this end, we develop protocols to address inner privacy based on secure logging.	To address this end, we develop protocols to address inner privacy based on secure logging.

6. Conclusion

Even after modifying the parameters several times, the best results we could obtain are not satisfying. Adding to that, setting up the T5 model with the use of Python, as a beginner, took a lot of time. Indeed, one of the main arguments in favor of machine simplification when opposed to manual simplification could be that it helps save a considerable amount of time to the person making the simplification. Moreover, any text automatically altered should be proofread and corrected before being published. In this particular case, the final automatic simplification was overall not very qualitative and contained a few meaning mistakes. It also was, in my opinion, under simplified and overall required as much if not more time for proofreading and correction as a manual simplification would have taken.

That said, many other tools can be used to automatically simplify any texts. Some of the most popular or easy to use are AI21 Studio and BERT. As for the T5 model, better results than those obtained for this task could probably be achieved with more time put into perfecting the settings.

7. Acknowledgements

This paper was created as part of the CLEF event of 2022 (Conference and Labs of the Evaluation Forum).

I would like to thank Liana Ermakova, from the University of Western Brittany, for her precious help in setting up the T5 model as well as for mentioning this CLEF event.

8. References

[1] Horacion Saggion, Automatic Text Simplification, Synthesis Lecture on Human Language Technologies, (2017).

[2] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu, (2019), Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

[3] Rémi Cardon, Approche lexicale de la simplification automatique de textes médicaux, UMR 8163 – STL – Savoirs Textes Langage, F-59000 Lille, France (2018).

[4] Natalia Grabar and Rémi Cardon. 2018. CLEAR – simple corpus for medical French. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 3–9, Tilburg, the Netherlands, November. Association for Computational Linguistics.

[5] Marco Alfano, Biagio Lenzitti, Giosuè Lo Bosco, Cinzia Muriana, Tommaso Piazza, Giovanni Vizzini, Design, development and validation of a system for automatic help to medical text understanding, *International Journal of Medical Informatics*, Volume 138, 2020, 104109, ISSN 1386-5056, https://doi.org/10.1016/j.ijmedinf.2020.104109.

[6] Ermakova, L., Bellot, P., Braslavski, P., Kamps, J., Mothe, J., Nurbakova, D., Ovchinnikova, I., SanJuan, E.: Overview of SimpleText 2021 - CLEF Workshop on Text Simplification for Scientific Information Access. In: Candan, K.S., Ionescu, B., Goeuriot, L., Larsen, B., Müller, H., Joly, A., Maistro, M., Piroi, F., Faggioli, G., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction, pp. 432–449, Lecture Notes in Computer Science, Springer International Publishing, Cham (2021), ISBN 978-3-030-85251-1

[7] Ermakova, L., Bellot, P., Braslavski, P., Kamps, J., Mothe, J., Nurbakova, D., Ovchinnikova, I., SanJuan, E.: Overview of SimpleText 2021 - CLEF Workshop on Text Simplification for Scientific Information Access. In: Candan, K.S., Ionescu, B., Goeuriot, L., Larsen, B., Müller, H., Joly, A., Maistro, M., Piroi, F., Faggioli, G., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction, pp. 432–449, Lecture Notes in Computer Science, Springer International Publishing, Cham (2021), ISBN 978-3-030-85251-1