Using a Pre-Trained SimpleT5 Model for Text Simplification in a Limited Corpus

José Monteiro [1], Micaela Aguiar [1] and Sílvia Araújo [1]

[1] *University of Minho, Braga, Portugal*

Abstract

In this paper, we describe our approach for solving Task 3 of the SimpleText Lab, organized as part of the Clef 2022 conference. The SimpleText Lab addresses issues of automatic text simplification of scientific texts in order to make scientific knowledge more accessible to everyone. To address Task 3, we trained Simple T5. In the first experiment, Simple T5 was trained with the small training dataset (648 entries) provided by the SimpleText team. Although there was a high number of unchanged sentences (49%), the sentences that were simplified and evaluated by SimpleText gathered the best overall result among the other participants. Nevertheless, we decided to run a new experiment training T5 with different datasets, namely, the original SimpleText training data set (trained with new parameters), WikiLarge, and a combination of WikiLarge and the SimpleText training dataset (WikiLast). We used EASSE metrics to compare the three models. Firstly, we tested the model with the TurkCorpus for reference and afterwards we tested them with the SimpleText Corpus. We wanted to know if a small but highly specialized dataset of the same discourse genre (abstracts) of the sentences that will be simplified combined with a larger general dataset would produce better results. WikiLast yielded the best SARI and BLEU results, however the BLEU results correlate with a higher percentage of exact matches. It seems that creating a small but highly specialized dataset may not make up for the investment, since the difference between the scores of the three models is not considerable.

Keywords

Text Simplification, SimpleT5, Sentence Transformers

1. Introduction

This paper describes an approach for building and designing a methodology for solving Task 3 of the SimpleText lab [1], organized as part of the Clef 2022 conference. The SimpleText 2022 Lab addressed text simplification approaches and proposed three taks: 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…) and TASK 3 Rewrite this! Given a query, simplify passages (sentences) from scientific abstracts. In this paper, we will propose a solution to address Task 3, using a pre-trained SimpleT5 model in the limited corpus of scientific abstracts.

2. Text Simplification

The first article on text simplification dates back to 1975 and was written in the field of social sciences [2]. Indeed, the fields of foreign and second language teaching have been interested in text simplification, as teachers need texts adapted to the level of their students and often end up having to create their own simplified materials. In the last two decades, (automated) text simplification has

[1]CLEF 2022 – Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy
EMAIL: jdiogoxmonteiro@gmail.com (A. 1); maguiar60@gmail.com (A. 2); saraujo@elach.uminho.pt (A. 3)
ORCID: 0000-0002-2904-3501 (A. 1); 0000-0002-5923-9257 (A. 2); 0000-0003-4321-4511 (A. 3)

become a field of natural language processing alongside matters of text summarization, machine translation and paraphrasing [2].

Text simplification relates to the process of modifying a text [3] — whether its syntax or lexicon or both —, in order to produce a simpler version of the text [4], that retains its original meaning [5], all the while improving its readability and understandability. Readability refers to the level of ease with which someone may read a text (it usually pertains to aspects such as grammar complexity or sentence length) [3]. Understandability (or comprehensibility) concerns the "amount of information a user may gain from a piece of text" [3] and it is influenced by factors such as the background knowledge the reader may have about a certain subject.

The applications of automated text simplification are many. Text simplification may be widely useful for second language learners [6], but also for assisted technology design to help people with Aphasia or lay readers faced with technical or specialized documents [3]. Automated text simplification can tackle, as well, the problem of scientific literacy [1]. The SimpleText Lab arises from the need to address the ever growing amount of scientific information people have to manage nowadays. Indeed, it is estimated that scientific knowledge doubles every five years [7]. Scientific texts are usually not easy to read and understand — they are complex and full of specialized terminology — since they are a product of a specialized discourse community mostly uninterested in science communication or popularization. Nowadays, it is more important than ever to make scientific texts more accessible to everyone.

3. Training SimpleT5 for Text Simplification

Text-to-Text transfer transformers (T5) are gaining popularity due to their competitive performance, effective scaling to larger model sizes, and ease of use in solving tasks as simple text-to-text mapping problems [8]. T5 is a model pre-trained in multiple NLP tasks, that proposes a unified approach, that is, the authors propose "to treat every text processing problem as a "text-to-text" problem, i.e. taking text as input and producing new text as output" [8]. T5 is trained on Colossal Clean Crawled Corpus (C4), a dataset created by applying a set of filters to English-language text sourced from the public Common Crawl web scrape. The T5 model is trained on several datasets for different tasks, such as Text Summarization, Question Answering Translation and Sentiment analysis [9]. To address Task 3, we face a series of constraints: lack of computing power, little development time and a fairly low amount of data available for pre-training the model. T5 is not trained for text simplification. However, after testing various models (mainly from the hugging face library [10], such as distilbart [11]), simpleT5 (base) seemed to give the best results with the least amount of training.

3.1. Initial Solution

In the first experiment, we trained SimpleT5 (base) with the data provided by the SimpleText Lab. The pre-training data consisted of a parallel corpus of 648 simplified sentences from Medicine and Computed Science scientific abstracts [1].

The sentences were simplified by experts and/or specialized translators. The test data set consisted of 116,763 sentences retrieved from the DBLP Citation Network Dataset.

The methodology we used was adapted from the process streamlined in this article [12] and the SimpleT5 documentation: 1. Pre-process the available training data and turn it into the acceptable by SimpleT5 format. The "simpletext_task3_train" was turned into our "source" and "simpletext_task3_decorated_run" into our "target"; 2. Split the full processed set between a training set and a testing set; 3. Train the model, using the simple methods given by SimpleT5; 4. Run every query in the "simpletext_task3_test" dataset through our model; 5. Evaluate the results; 6. If the results don't meet our standards: Go back to step 3. with updated arguments for the model; Else: Go to next step; 7. Build a script that pushes every result into the file t format required by the SimpleText Lab. This was certified using the available script; 8. Submit the Results.

The final parameters for the model were the following: source_max_token_len = 128, target_max_token_len = 50, batch_size = 8, max_epochs = 5. The actual training dataset was composed of 80% of the original training set (518 entries) and the other entries (130) were needed for validation.

The general results showed that from the total 116,763 sentences our model didn't change 42,189 sentences (49%) and 852 (0,72%) sentences were truncated. 3,217 sentences became longer (2,7%) and the syntax complexity (2.94) and the lexical complexity (3.06) scores were average when comparing to the total of runs submitted: the highest score for syntax complexity was 4.69 and the lowest 2.10 and the highest score for lexical complexity was 3.69 and the lowest 2.42. Regarding informational loss, our run had the lowest score 1.50, the highest being 3.84. 564 sentences were evaluated for information distortion: there were 9 instances of non-sense (1,5%), 3 instances of contresens (0,5%), 4 instances of topic shifts (0,70%), 3 instances of wrong synonyms, 19 instances of ambiguity (3,3%), 94 instances of omission of essential details (16,6%), 9 instances of overgeneralization (1,5%), 13 instances of oversimplification (2,3%), 2 instances of unsupported information (0,3%) and 2 of unnecessary details (0,3%), 5 instances of redundancy (0,8%) and 1 instances of style (0,1%). In the final ranking for Task 3, we placed first with a score of 0.149 (the second and third places score respectively 0.122 and 0.119).

3.2. Training SimpleT5 with Different Datasets

Despite our classification, we decided to run a second experiment training T5 with different datasets. In order to evaluate/test the new experiments, we used the Python package EASSE [13] that offers a single-point access to popular automatic metrics for sentence simplification, such as BLEU, SARI and FKGL. BLEU is used to assess grammaticality and meaning preservation; SARI examines simplicity gain (word added, deleted and kept) and FKGL indicates sentence length and number of syllables. Additionally, EASSE makes available three datasets for automatic sentence simplification evaluation: PWKP, TurkCorpus and HSplit. EASSE also offers quality estimation metrics (the compression ratio, Levenshtein similarity, average number of sentence splits, proportion of exact matches, proportion of added words, deleted words, and lexical complexity score) [13].

We decided to train T5 with three different datasets. The first model (New SimpleText) uses, as the name suggests, the original SimpleText dataset. This time we opted to follow the most popular state of the art models both in terms of parameters and validation sets. That being said, for training we used the original set with all the entries (648), and for validation TurkCorpus was used (2000 entries). The validation set available from TurkCorpus and the final parameters were the same for all three models: source_max_token_len = 256, target_max_token_len = 128, batch_size = 8, max_epochs = 5, precision = 32. The second model (WikiLarge) used the WikiLarge dataset [14]. WikiLarge is a popular dataset used for text simplification: it contains 296,402 sentence pairs from English Wikipedia and Simple Wikipedia. The third and final model (WikiLast) was trained with WikiLarge + SimpleText Training Dataset. We wanted to know if a small but highly specialized dataset of the same discourse genre (abstracts) of the sentences that will be simplified combined with a larger general dataset would gather different and better results. If our tests reveal that the merger of a large generic set and a small specialized one for model training have a significant increase in the model accuracy, we could make the case that focusing resources on building small specialized sets could be a good investment.

Having the three final models trained and ready, we tested them with TurkCorpus, in order to have benchmark values. Table 1 shows the EASSE metrics results of the three models. The New SimpleText Model scores the highest value of BLEU (94.897), however as Sulem, Abend & Rappoport point out, BLEU "gives high scores to sentences that are close or even identical to the input" [15]. Indeed, the New SimpleText Model has a high value of exact copies (47%). Between the WikiLarge and the WikiLast Models the differences do not appear to be significant. It is noteworthy that the WikiLast gathered the best SARI result.

Table 1
EASSE metrics: Models Tested with the TurkCorpus

Metrics	BLEU	SARI	FKGL	Exact Cop.	Lexic. Comp.
New SimpleT	94.897	35.127	9.586	0.479	8.268
WikiLarge	86.706	37.652	9.17	0.295	8.18
WikiLast	88.063	38.165	9.056	0.312	8.186

Table 2 shows an example taken from the TurkCorpus and simplified by the models. The major transformation in this sentence was the substitution of the word "victorious" with a simpler word: New SimpleText replaced it with "elected" and WikiLarge and WikiLast replaced it with "won". Arguably, "elected" is less complex and more common than "victorious", but more complex and less common than "won". This example illustrates well the lexical complexity scores from Table 1.

Table 2
Example of Simplified Sentence of Models Tested with TurkCorpus

Original	In 1998, Culver ran for Iowa Secretary of State and was victorious.
New SimpleT	Culver ran for Iowa Secretary of State in 1998 and was elected.
WikiLarge	In 1998, Culver ran for Iowa Secretary of State and won.
WikiLast	In 1998, Culver ran for Iowa Secretary of State and won.

Next, the models were tested with a selection of the original SimpleText dataset (10,000 entries). Table 3 shows the EASSE metrics results of the three models tested with the SimpleText dataset. In comparison with the other two models, WikiLast produces the best SARI results. Wikilast also presented the best BLEU results, even though they correlate with an increase in exact matches (54%). In this run, the NewSimpleText Model and WikiLarge are very similar in most metrics.

Table 3
EASSE metrics: Models Tested with the SimpleText Corpus

Metrics	BLEU	SARI	FKGL	Exact Cop.	Lexic. Comp.
New SimpleT	81.084	29.847	13.469	0.374	8.794
WikiLarge	83.915	30.395	14.067	0.478	8.778
WikiLast	88.511	31.269	13.949	0.542	8.787

Table 4 and Table 5 show examples taken from the three models. In the first example (Table 4), in the New SimpleText Model the sentence remains unchanged. We can also see that both WikiLarge and WikiLast unabbreviate the verb form, but only WikiLast replaces the verb form "developed" with a simple and more common one ("made").

Table 4
Example 1 of Simplified Sentence of Models Tested with SimpleText Corpus

Original	It's developed for Windows Mobile® for use in mobile devices such as PDA's and Smartphone's.
New SimpleT	It's developed for Windows Mobile® for use in mobile devices such as PDA's and Smartphone's.
WikiLarge	It was developed for Windows Mobile® for use in mobile devices such as PDAs and Smartphones.
WikiLast	It was made for Windows Mobile® for use in mobile devices such as PDAs and Smartphones.

In Table 5, the sentence remains unchanged in the New Simple Text and WikiLarge models, while WikiLast simplifies the sentence, by replacing "browse" with "use".

Table 5
Example 2 of Simplified Sentence of Models Tested with SimpleText Corpus

Original	The BioWAP service can be browsed with any WAP terminal.
New SimpleT	The BioWAP service can be browsed with any WAP terminal.
WikiLarge	The BioWAP service can be browsed with any WAP terminal.
WikiLast	The BioWAP service can be used with any WAP terminal.

4. Discussion and Conclusion

Unsurprisingly, when the models were tested with the TurkCorpus, the New Simple Text model yielded the worst SARI results, since we are using a model trained on specialized data to simplify more generic data. WikiLast produced the best SARI result, which we think could be simply attributed to its larger data set. In the TurkCorpus case, just as we expected, the difference in results between WikiLarge and WikiLast are too small and unnoticeable to justify the need for the specialized set.

When the models were tested with the SimpleText Corpus, the New Simple Text model yielded the worst SARI results of both runs. WikiLast had the best SARI and BLEU results, however, as mentioned before, the high BLEU results correlate with a higher percentage of exact copies. These results have the limitation of stemming from automatic metric. As Shardlow observes, automatic metrics are largely ineffective [3], and they are not a substitute for human judgment. Nevertheless, if we take into account the results, it seems that creating a small but highly specialized dataset may not make up for the investment, since the difference between the scores of the three models is not considerable. Evidently, it is possible that if the specialized dataset was larger the results would be better and that larger specialized datasets (of scientific texts, judicial texts or academic texts, for example) could produce better results than generic datasets.

5. References

[1] E. Liana, P. Bellot, J. Kamps, D. Nurbakova, I. Ovchinnikova, E. SanJuan, E. Mathurin, S. Araújo, R. Hannachi, and S. Huet, Overview of the CLEF 2022 SimpleText Lab: Automatic Simplification of Scientific Texts, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Thirteenth International Conference of the CLEF Association (CLEF 2022), edited by Alberto Barrón-Cedeño, Giovanni Da San Martino, Mirko Degli Esposti, Fabrizio Sebastiani, Craig Macdonald, Gabriella Pasi, Allan Hanbury, Martin Potthast, Guglielmo Faggioli, and Nicola Ferro, 27, 2022.

[2] H. Z. Özcan, Z. Batur, A Bibliometric Analysis of Articles on Text Simplification: Sample of Scopus Database, International Journal of Education and Literacy Studies 9 (2) 24 (2021). https://doi.org/10.7575/aiac.ijels.v.9n.2p.24.

[3] M. Shardlow. A Survey of Automated Text Simplification, International Journal of Advanced Computer Science and Applications 4 (1) (2014). https://doi.org/10.14569/SpecialIssue.2014.040109.

[4] M. Louis, S. Humeau, P.-E. Mazaré, A. Bordes, É. V. de La Clergerie, and B. Sagot, Reference-Less Quality Estimation of Text Simplification Systems, arXiv (2019). http://arxiv.org/abs/1901.10746.

[5] Š. Sanja, M. Franco-Salvador, S. P. Ponzetto, P. Rosso, and H. Stuckenschmidt. Sentence Alignment Methods for Improving Text Simplification Systems, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pp. 97–102. Vancouver, Canada, 2017, pp. 97–102. https://doi.org/10.18653/v1/P17-2016.

[6] X. Menglin, E. Kochmar, and T. Briscoe, Text Readability Assessment for Second Language Learners, in: Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, 2016, pp. 12–22. https://doi.org/10.18653/v1/W16-0502.

[7] J. Cribb, T. Sari, Open Science: Sharing Knowledge in the Global Century. Victoria, Collingwood, 2010.

[8] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv, (2020). URL: http://arxiv.org/abs/1910.10683.

[9] H. Digaari, Solved! Google's Text-To-Text Transfer Transformer (T5) Bottleneck, Towards Data Science, 2021. URL: https://bit.ly/3NAp24W

[10] Hugging Face, URL: https://huggingface.co/

[11] S. Shleifer, distilbart-cnn-6-6. URL: https://huggingface.co/sshleifer/distilbart-cnn-6-6

[12] S. Roy, SimpleT5 - Train T5 Models in Just 3 Lines of Code, 2021. URL: https://bit.ly/38MS7eZ

[13] F. Alva-Manchego, L. Martin, C. Scarton, L. Specia, 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,Association for Computational Linguistics, 2019, Hong Kong, China, pp. 49–54. https://doi.org/10.18653/v1/D19-3009.
[14] Z. Xingxing, M. Lapata. Sentence Simplification with Deep Reinforcement Learning, arXiv (2017). http://arxiv.org/abs/1703.10931.
[15] S. Elior, O. Abend, and A. Rappoport, Semantic Structural Evaluation for Text Simplification, arXiv (2018). http://arxiv.org/abs/1810.05022