Controllable Sentence Simplification Using Transfer Learning

Antonio Menta [a] and Ana Garcia-Serrano [a]

[a] *E.T.S.I. Informática (UNED), Spain*

Abstract

The CLEF 2022 SimpleText track addresses the challenges of text simplification approaches to generate versions of scientific texts for a non-expert audience in highly technical domains, such as Computer Science or Medicine. Our work uses the transfer learning capabilities of the T5 pre-trained language model, adding a method to control specific simplification features. We present a new feature based on masked tokens prediction (Language Model Fill-Mask) to control the lexical complexity of the text generation process. The results obtained with the SARI metric are at the same level as previous work in other domains for sentence simplification.

Keywords [1]

Controllable features, sentence simplification, language models

1. Introduction

In recent decades, the volume of information from multiple sources has grown exponentially. This means that, on many occasions, a given person receives textual information that is not adapted to their level of comprehension, making it difficult to understand and even generating rejection of the subject matter. This may be due to the use of specific vocabularies in areas such as Computer Science, Finance or Medicine [1], where neologisms and recently created words appear. It also occurs with cognitive disabilities, such as dyslexia or aphasia, where certain words or syntactic structures may be difficult to understand. Digital Humanities (DH) [2] is a particularly complicated area, where language has evolved over the centuries, leading to the disuse of certain words and syntactic structures. Moreover, many of them have changed their meaning, acquiring a completely different one from the original. The use of automatic text simplification methods can improve reading comprehension and scientific dissemination.

Lexical Simplification (LS) is the process of replacing complex words in a given sentence with simpler alternatives of equivalent meaning [3]. Syntactic simplification attempts to reduce grammatical complexity by finding the most complex syntactic structures and replacing them with simpler ones. Finally, sentence simplification aims to reduce the complexity of a sentence while preserving the meaning and important information.

Recent efforts in sentence simplification have been influenced by the techniques used for machine translation and summarization. The simplification task is often treated as monolingual translation, where a complex sentence is translated into a simpler one, although with a different goal. In a summary, the aim is a reduced version of a text, where the most important thing is usually the size of the text obtained, keeping as much information as possible. In a machine translation task, the aim is to retain the information regardless of the final size of the text, while simplification aims to make the content easier to understand. This usually leads to a compressed version of the text, but it is not mandatory. The explanation of a concept or acronym may result in a longer text than the original.

[1] CLEF 2022 – Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy
EMAIL: amenta1@alumno.uned.es (Antonio Menta); agarcia@lsi.uned.es (Ana Garcia-Serrano)
ORCID: 0000-0002-3172-2829 (Antonio Menta); 0000-0003-0975-7205 (Ana Garcia-Serrano)

CEUR Workshop Proceedings (CEUR-WS.org) Proceedings

2. Related work

Currently, there has been a shift towards work based on deep learning techniques for sentence simplification, similar to what has happened with most natural language processing (NLP) tasks. One line of research is the use of sequence-to-sequence-based neural networks to predict explicit edit operations (adding, deleting and keeping) to transform the original sentence into a simpler version. A neural Programmer-Interpreter approach, inspired by the way humans perform iterative simplification is proposed by Dong et al.[4]. Later, Cumbicus-Pineda et al. [5], added a convolutional graph module with the syntactic information of sentences to aid the detection of complex structures in the simplification process.

Deep learning language models, in particular, *pretrained Bidirectional Encoder Representations from Transformers* (BERT) [6] and transfer learning techniques since ULMFit [7], have revolutionized sentence simplification task approaches. Nowadays, it is not necessary to train a model from scratch and there is a tendency to start from pre-trained language models such as BERT, RoBERTa [8] or GPT-3 [9]. Within this type of model, those based on an encoder-decoder architecture have excelled in tasks such as text summarization or machine translation. Among them, BART [10], T5 [11] and mT5 [12], the latter for texts in a language other than English, are particularly noteworthy.

3. SimpleText laboratory at CLEF 2022

The SimpleText laboratory at CLEF 2022 addresses the study of different approaches for text simplification to promote access to scientific information. One of its main purposes is the creation of a community of researchers interested in such a complex task in the world of NLP as the simplification of scientific texts [13]. The information shared in the laboratory focuses on two fields with many technical terms, Medicine and Computer Science. In these fields, when a simplified version of the content of an article is created, it is done from the perspective of an expert in the field, so the result is usually a shortened version of the article full of technicalities. This makes it difficult to be understood by a non-expert audience.

This laboratory is the continuation of the SimpleText workshop, proposed last year for CLEF 2021 [14]. One of the most important parts of the workshop was to search for possible answers to several questions and doubts about text simplification: which terms should be simplified, which terms should be contextualized by giving a definition and/or application, how to improve the readability of a given short text without losing information, or which metrics are appropriate to evaluate the task.

3.1. Datasets

The organization provided a dataset for each of the laboratory tasks. The source of the data to be simplified was the *Citation Network Dataset: DBLP+Citation, ACM Citation network (12th version)*. For documents belonging to the Computer Science domain, the organization selected 13 topics based on headlines from *The Guardian* newspaper. Medical papers were obtained from *Google Scholar* and *Pubmed* articles. The text passages extracted from the abstracts were simplified by a master's student in translation or by an expert, either a subject matter expert or a professional translator. Each example was rewritten several times until it was clear to a non-expert in the field.

For task 3, the examples have several identifiers (document, query, and sentence), the original text, the query text and, in a separate file, the simplified text. The training and test dataset consists of 648 and 116763 instances respectively.

4. CLARA-HD model for Text Simplification

Deep Learning-based language models have proven to be state-of-the-art in multiple NLP tasks. Its effectiveness comes from prior training on a self-supervised task, such as a language model or a masked token prediction process. Once the model is trained, its use in other tasks requires much less labelled text than training a model from scratch. Among all types of models, T5 has shown promising results in different tasks such as text summarization, question-answering and classification problems [11].

We selected a transfer learning approach because the training dataset has only 648 instances. Due to the small size of the training data, we trained with a larger dataset, WikiLarge [15]. This is one of the largest and most widely used datasets for text simplification. It consists of 296,402 automatically aligned sentence pairs extracted from the English version of Wikipedia and Simple Wikipedia [16], [17], [18]. We left the shared task dataset to validate the trained model and find the best control token hyperparameters with it.

In addition, we explored the combination of using a pre-trained language model, with a token control mechanism based on different features to control the simplification level, similar to the work done by Sheang and Saggion [19].

4.1. Features used for control tokens

In our work, we have used control tokens to modify several features of the text that previous approaches have shown to be useful for text simplification. These include the amount of text compression (Chars), word length (Words), paraphrasing (LevSim) and syntactic complexity (DepTreeDepth) [19], [20].

In our case, we added a feature based on the lexical complexity of the words (Fill-Mask). This feature uses the masking of certain tokens of sentences for subsequent prediction, under the hypothesis that the position within the prediction ranking of the words in the simple version will be lower than in its complex version. If a sentence contains simple concepts, the Fill-Mask model will be able to predict them before the complex version.

To predict masked tokens, we used the COVID-SciBERT model [21], an expanded version of SciBert [22] with the articles present in the competition COVID-19 Open Research Dataset Challenge (CORD-19)[2]. Therefore, the model has been trained in the two domains, Computer Science and Medicine. Due to the high computation times, masked tokens have been only nouns, verbs and adjectives, one at a time. An example can be found in Table 1. Once predictions are obtained, the position of the masked words within the ranking is stored, and finally, each sentence is identified by the median of the results obtained.

Table 1
Fill-Mask example

Model	Text
Before Masking	Based on the inception - v3 architecture, our system performs better in terms of processing complexity and accuracy than many existing models for imitation learning.
COVID-SciBERT	Based on the [MASK] - v3 architecture, our system performs better in terms of processing complexity and accuracy than many existing models for imitation learning.

In summary, the text features used for the simplification check were:

- Chars (CLR): Character length ratio between the original and the target sentence (the simplified version). The number of characters in the simplified version is divided by the number of characters in the original. Previous work has shown a correlation between simplicity and the number of characters in the sentence [23].
- LevSim (LR): Levenshtein normalized similarity at character-level [16] between the original and the simplified version. This feature is a measure of the modifications made including its paraphrase level.

[2] https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge

- DepTreeDepth (DTDR): Maximum depth of the dependency tree of the simplified version divided by that of the source, under the assumption that a simple sentence makes use of syntactic structures with fewer dependencies than its complex version.
- Words (WLR): Ratio of the number of words between the original and the simplified sentence. The number of words in a simple sentence is divided by the number of words in the complex sentence.
- Language Model Fill-Mask (LMFMR): Position within the prediction ranking of all masked words in the simplified version divided by that of the original. A language model trained on a masking task can predict earlier the set of masked words in a simple sentence than in a complex sentence. LMFMR feature values distribution of the training dataset can be found in Figure 1.

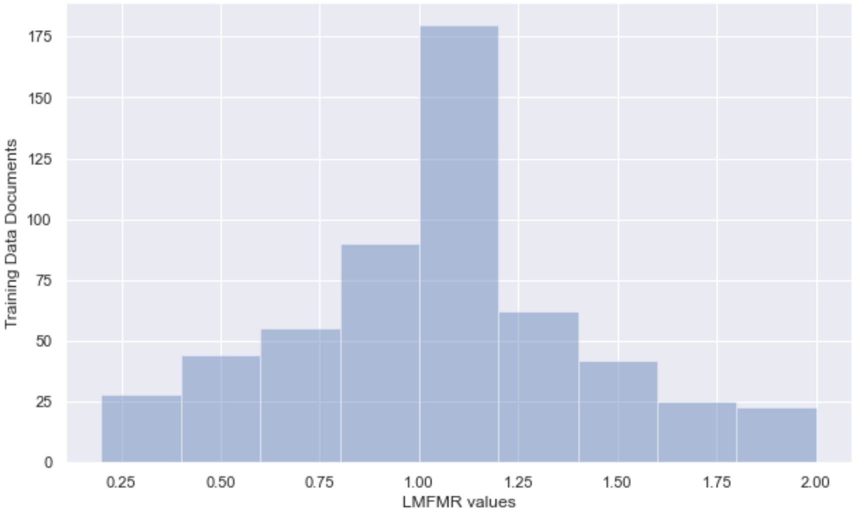

Figure 1: LMFMR feature values distribution

During training, for each pair of sentences, original and simple, the features are calculated and then the value of the simple version is divided by the original one. This determines a value for each pair that is added as a prefix to the original text, as it can be found in Table 2 (being preceded by the feature identifier).

Table 2
Modified example with calculated features

Example	Text
Original	In the modern era of automation and robotics, autonomous vehicles are currently the focus of academic and industrial research.
Modified	simplify: CLR_0.65 DTDR_0.75 LMFMR_0.65 LR_0.65 WLR_0.8 In the modern era of automation and robotics, autonomous vehicles are currently the focus of academic and industrial research.

At inference time, the values of each feature can be modified to control the sentence compression. T5 is a Text-to-Text model in which both the input and output are text. For this purpose, the prefix "simplify:" is added to the input text. This format enables different combinations of features without changing the architecture of the model. Furthermore, thanks to this format, the testing bench becomes faster because the values can be calculated in advance, stored and added at the time of training.

4.2. Experiments

Due to the large size of current language models, we trained with the T5-small version (60 million parameters) on an Nvidia GeForce GTX 1070 Ti GPU, with 8GB of memory. The model was trained using the Pytorch-lightning and HuggingFace libraries [24]. All tests have been performed with the same hyperparameters such as 5 epochs, a batch size of 6 for both training and validation, a maximum number of 256 tokens, a learning rate of 3e-4 and a weight_decay of 0.1. The training time was 36 hours for feature generation, mainly due to the long prediction time involved in the Fill-Mask feature, and 20 hours for training. The total time was 56 hours.

Once the model was trained, a hyperparameter search was performed using Optuna [23]. Five hundred experiments were performed, limiting each feature's maximum and minimum values to 1 and 0.3, respectively. A value of 0.05 was chosen as the incremental value for the search. Several examples are shown in table 3.

Table 3
Hyperparameter search examples

SARI	WLR	CLR	LR	DTDR	LMFMR
37.40	0.75	0.6	0.6	0.95	0.75
36.34	0.7	0.6	0.5	0.5	0.8
35.33	0.7	0.7	0.7	0.8	0.75
33.38	0.7	1	1	0.7	0.9
32.51	0.4	0.6	0.5	0.7	0.3
31.94	0.3	0.8	0.8	0.5	0.5

4.3. Metric

Following previous research in recent years [25], [26], [4], we have selected SARI as a metric [27] to validate our results. SARI compares the system output with the references and the input sentence. It averages F1-scores for adding, keeping and deleting operations. We calculate SARI with the EASSE simplification evaluation suite [28], which fixes several bugs and inconsistencies found in the original version. The main differences are[3]:

1. EASSE applies the same normalization to source, prediction and references. The original implementation only on the prediction and reference.
2. The original SARI implementation tokenizes the input text twice. This causes differences between the tokenization of the training and test sets.
3. The original implementation had an overflow bug when ngram statistics exceeded the maximum limit for integers.

5. Results

From the hyperparameter search performed with Optuna, the best result was obtained with CLR=0.6, DTDR=0.95, LMFMR=0.75, LR=0.6 and WLR=0.75. With these hyperparameters, the final result was a SARI value of 37.40. Adding our feature, the model obtains a +0.35 SARI improvement. The importance of the features is shown in table 4. It is important to highlight that, although the model has not been trained with specific examples from the domains of the task, Computer Science and Medicine, the results obtained with the SARI metric have similar values to other approaches trained in different domains.

[3] https://github.com/feralvam/easse

For example, the SOTA[4] on Newsela dataset [29] is obtained by a transformer-based seq2seq model [30] with a SARI value of 36.6. Another example is the TurkCorpus dataset [27], where a multilingual unsupervised sentence simplification system [25] using sentence-level paraphrase data instead of proper simplification data, reported a SARI value of 42.62.

Table 4
Features importance in final results

SARI	WLR	CLR	LR	DTDR	LMFMR
37.40	0.75	0.6	0.6	0.95	0.75
37.05	0.75	0.6	0.6	0.95	--
36.89	0.75	0.6	0.6	--	0.75
35.16	0.75	0.6	--	0.95	0.75
36.06	0.75	--	0.6	0.95	0.75
35.80	--	0.6	0.6	0.95	0.75
24.26	0.75	--	--	--	--
25.99	--	0.6	--	--	--
25.25	--	--	0.6	--	--
23.61	--	--	--	0.95	--
23.66	--	--	--	--	0.75

6. Conclusion and future work

In our work[5], we propose the use of a pre-trained Deep Learning language model in a simplification task using its transfer learning capabilities. A domain-general dataset such as Wikipedia has been used to train the model and the proposed dataset has been used for validation. Because of the calculated features: text compression, word length, lexical and syntactic complexity, and the level of paraphrasing, the model has been able to simplify and obtain similar results to previous work, even without being trained directly on the domain data.

Our approach enables us to control the simplification result by selecting specific values for each of the features previously trained. This increases the flexibility of the system, as it is possible to generate different simplified versions, controlling e.g., the length of the text or its lexical richness. Furthermore, the architecture supports new features without having to modify any previous components, making it an ideal approach for discovering new ones in the future.

We plan to further explore in depth the capabilities of our contribution, the Fill-Mask feature. For the shared task, we have had little time to explore its possibilities. We are also planning to work with more powerful hardware to test bigger models.

7. Acknowledgements

This work has been partially funded by the CLARA-NLP coordinated project with subprojects in three domains: History [31] (PID2020-116001RB-C32), Biomedicine [32] (PID2020-116001RA-C33) and Economics [33] (PID2020-116001RB-C31). The CLARA-NLP project integrates the experience of researchers from three institutions, UAM, UNED and CSIC, who have collaborated for more than a decade on different NLP projects. On this occasion, we explore the applications of automatic simplification of speciality text into easier-to-read text. Therefore, the common basis will be the evaluation framework and the difference will be in the domain (Financial, Digital Humanities and Medical) where each group has a contrasting experience in previous projects.

[4] https://paperswithcode.com/sota/text-simplification-on-newsela

[5] Code is released with an open-source license at https://github.com/Hisarlik/simpleTextCLEF

8. References

[1] L. Campillos-Llanos, A. Valverde-Mateos, A. Capllonch-Carrión, and A. Moreno-Sandoval, 'A clinical trials corpus annotated with UMLS entities to enhance the access to evidence-based medicine', BMC Medical Informatics and Decision Making., vol. 21, no. 1, 2021 pp. 1–19.

[2] A. Garcia-Serrano and A. Menta, 'La inteligencia artificial en las humanidades digitales: un caso de éxito y un caso de estudio en un corpus digital de historia del arte', La Revista de Humanidades Digitales, vol. 7, 2022.

[3] G. H. Paetzold and L. Specia, 'A survey of lexical simplification', Journal of Artificial Intelligence Research, vol. 60, 2017, pp. 549–593.

[4] Y. Dong, Z. Li, M. Rezagholizadeh, and J. C. K. Cheung, 'Editnts: An neural programmer-interpreter model for sentence simplification through explicit editing', in ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 2020, pp. 3393–3402.

[5] O. M. Cumbicus-Pineda, I. Gonzalez-Dios, and A. Soroa, 'A Syntax-Aware Edit-based System for Text Simplification', International Conference Recent Advances in Natural Language Processing, RANLP, 2021, pp. 324–334.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding', 2018.

[7] J. Howard and S. Ruder, 'Universal language model fine-tuning for text classification', in ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers), 2018, vol. 1, pp. 328–339.

[8] Y. Liu et al., 'RoBERTa: A Robustly Optimized BERT Pretraining Approach', Jul. 2019.

[9] T. B. Brown et al., 'Language models are few-shot learners', Advances in Neural Information Processing Systems, 2020.

[10] M. Lewis et al., 'BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension', 2020, pp. 7871–7880.

[11] C. Raffel et al., 'Exploring the limits of transfer learning with a unified text-to-text transformer', Journal of Machine Learning Research., vol. 21, 2020, pp. 1–67.

[12] L. Xue et al., 'mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer', in NAACL, 2021, pp. 483–498.

[13] L. Ermakova, P. Bellot, J. Kamps, D. Nurbakova, I. Ovchinnikova, E. SanJuan, E. Mathurin, S. Araújo, R. Hannachi, S. Huet and N. Poinsu. 'Automatic Simplification of Scientific Texts: SimpleText Lab at CLEF-2022', M. Hagen, S. Verberne, C. Macdonald, C. Seifert, K. Balog, K. Nørvåg, V. Setty (Eds.), Advances in Information Retrieval, vol. 13186, 2022, pp. 364–373. Springer International Publishing. https://doi.org/10.1007/978-3-030-99739-7_46.

[14] L. Ermakova, P. Bellot, P. Braslavski, J. Kamps, J. Mothe, D. Nurbakova, I. Ovchinnikova, E. SanJuan. 'Overview of SimpleText CLEF 2021 Workshop and Pilot Tasks', 2021.

[15] X. Zhang and M. Lapata, 'Sentence simplification with deep reinforcement learning', in EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings, 2017, pp. 584–594.

[16] D. Kauchak, 'Improving text simplification language modeling using unsimplified text data', in ACL 2013 - 51st Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 2013, vol. 1, pp. 1537–1546.

[17] K. Woodsend and M. Lapata, 'Learning to simplify sentences with quasi-synchronous grammar and integer programming', in EMNLP 2011 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 2011, pp. 409–420.

[18] Z. Zhu, D. Bernhard, and I. Gurevych, 'A monolingual tree-based translation model for sentence simplification', in Coling 2010 - 23rd International Conference on Computational Linguistics, Proceedings of the Conference, 2010, vol. 2, pp. 1353–1361.

[19] K. C. Sheang and H. Saggion, 'Controllable Sentence Simplification with a Unified Text-to-Text Transfer Transformer', INLG 2021 - 14th International Conference on Natural Language Generation, Proceedings, 2021 pp. 341–352.

[20] L. Martin, É. V. de la Clergerie, B. Sagot, and A. Bordes, 'Controllable sentence simplification',

LREC. 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings, 2020, pp. 4689–4698.

[21] Tanmay Thakur, ‘COVID-SciBERT: A small language modelling expansion of SciBERT, a BERT model trained on scientific text.’, 2021. [Online]. Available: https://huggingface.co/lordtt13/COVID-SciBERT.

[22] I. Beltagy, K. Lo, and A. Cohan, ‘SCIBERT: A pretrained language model for scientific text’, EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference. pp. 3615–3620, 2019.

[23] L. Martin, S. Humeau, P. E. Mazare, A. Bordes, E. D. La Clergerie, and S. Benoit, ‘Reference-less Quality Estimation of Text Simplification Systems’, ATA 2018 - 1st Workshop on Automatic Text Adaptation, Proceedings of the Workshop, pp. 29–38, 2018.

[24] T. Wolf et al., ‘HuggingFace’s Transformers: State-of-the-art Natural Language Processing’, ArXiv, vol. abs/1910.0, 2019.

[25] L. Martin, A. Fan, É. de la Clergerie, A. Bordes, and B. Sagot, ‘MUSS: Multilingual Unsupervised Sentence Simplification by Mining Paraphrases’, arXiv, 2020.

[26] A. Awasthi, S. Sarawagi, R. Goyal, S. Ghosh, and V. Piratla, ‘Parallel iterative edit models for local sequence transduction’, EMNLP-IJCNLP 2019- Conference on Empirical Methods in Natural Language Processing 9th International Joint Conference on Natural Language Processing. Proceedings of the Conference, pp. 4260–4270, 2020.

[27] W. Xu, C. Napoles, E. Pavlick, Q. Chen, and C. Callison-Burch, ‘Optimizing Statistical Machine Translation for Text Simplification’, Transactions of the Association for Computational Linguistics, vol. 4, 2016, pp. 401–415.

[28] F. Alva-Manchego, C. Scarton, L. Martin, and L. Specia, ‘EASSE: Easier automatic sentence simplification evaluation’, in EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Proceedings of System Demonstrations, 2019, pp. 49–54.

[29] W. Xu, C. Callison-Burch, and C. Napoles, ‘Problems in Current Text Simplification Research: New Data Can Help’, Transactions of the Association for Computational Linguistcs., vol. 3, pp. 283–297, 2015.

[30] C. Jiang, M. Maddela, W. Lan, Y. Zhong, and W. Xu, ‘Neural CRF Model for Sentence Alignment in Text Simplification’, in ACL, 2020.

[31] A. Menta, E. Sanchez-Salido, and A. García-Serrano, ‘Transcripción de periódicos históricos: aproximación CLARA-HD’, Procesamiento del Lenguaje Natural, vol. 69, 2022.

[32] L. Campillos-Llanos, A. Terroba, S. Zakhir, A. Valverde, and A. Capllonch, ‘Building a comparable corpus and a benchmark for Spanish medical text simplification’, Procesamiento del Lenguaje Natural, vol. 69, 2022.

[33] A. Moreno-Sandoval, A. Gisbert, and H. Montoro, ‘FinT-esp : A corpus of financial reports in Spanish’, Multiperspectives in Analysis and Corpus Design, Granada, Editorial Comares, 2020, pp. 89–102.