Text Simplification with Autoregressive Models

Alena Fenogenova
Sberbank, SberDevices
Moscow, Russia
alenush93@gmail.com

Abstract

Text Simplification is the task of reducing the complexity of the vocabulary and sentence structure of the text while retaining its original meaning with the goal of improving readability and understanding. We explore the capability of the autoregressive models such as RuGPT3 (Generative Pre-trained Transformer 3 for Russian) to generate high quality simplified sentences. Within the shared task RuSimpleSentEval we present our solution based on different usages of RuGPT3 models. The following setups are described: 1) few-shot unsupervised generation with the RuGPTs models 2) the effect of the size of the training dataset on the downstream performance of fine-tuned model 3) 3 inference strategies 4) the downstream transfer and post-processing procedure using pre-trained paraphrasers for Russian. This paper presents the second-place solution on the public leaderboard and the fifth-place solution on the private leaderboard. The proposed method is comparable with the novel state-of-the-art approaches. Additionally, we analyze the performance and discuss the flaws of RuGPTs generation.

Keywords: text simplification, RuGPT3, text generation, paraphrase generation
DOI: 10.28995/2075-7182-2021-20-227-234

Упрощение текстов с помощью авторегрессионных моделей

Алена Феногенова
Сбербанк, SberDevices
Москва, Россия
alenush93@gmail.com

Аннотация

Упрощение текста — задача автоматического получения упрощенного предложения из сложного. В работе представлена методика упрощения текстов на основе авторегрессионных моделей, в частности RuGPT3 (Generative Transformer 3 for Russian). Решение представлено в рамках соревнования RuSimpleSentEval, которое заняло второе место на публичном лидерборде по метрике SARI и пятое место на приватном лидерборде. В работе рассмотрены следующие подходы: 1) генерация упрощенного текста с помощью техники few-shot, 2) изучение влияния размера обучающей выборки и параметров на целевое качество моделей, обученных с помощью метода fine-tuning, 3) сравнение трех инференс стратегий и постборaботки, 4) применение предобученных моделей для генерации парафразов на русском языке на целевой задаче и в качестве компонента пост-обработки сгенерированных упрощенных текстов.

Ключевые слова упрощение текстов, RuGPT3, генерация парафразов

1 Introduction

The task of text simplification (TS) aims to reduce its linguistic complexity in order to improve readability and understanding. Text complexity criteria include the presence of complex grammatical structures, participial and adverbial constructions, subordinate sentences, the presence of infrequent and ambiguous words. Recent research on TS has been of keen interest, especially after the development of automatic approaches which have led to the transition from manually defined rules to automatic simplification using neural networks. Simplification has a variety of important applications. For example, in socio-psychological respect, it increases the information accessibility for those with cognitive disorders such as aphasia, dyslexia, and autism, as well as for non-native speakers. Furthermore, automatic text simplification could improve performance on other NLP tasks, such as paraphrasing, summarization, information extraction, semantic role labeling, and machine translation.

Existing methods have been predominantly designed for English due to the availability of high-quality text corpora which contain aligned complex and simplified sentences such as Newsela[1] [24] and Turk Corpus [25]. WikiLarge constructed from Wikipedia and Simple Wikipedia is a very common dataset for English as well. However, the construction of such datasets for new language is expensive, and no attempts have been made to create a TS dataset for the Russian language. To this end, the shared task RuSimpleSentEval-2021 [16] aims to fill this gap and facilitate the development of automatic TS methods for Russian. This paper describes the submission to the shared task and proposes the TS method based on RuGPT3, and details the experiments with the autoregressive models for Russian. We explore the RuGPT3[2] models capabilities in a full compliance with the competition rules, study the effect of the size of the training dataset on the model performance, combine different inference strategies and post-processing techniques. The method has achieved the second place on the RuSimpleSentEval public leaderboard and the fifth place on the private leaderboard.

The remainder is organized as follows: Section 2 briefly describes the prior research in the field; Section 3 outlines the data used in the experiments; Section 4 provides the description of the experiments; we discuss the results and provide the analysis of the proposed method and generated abilities of the best model in Section 5, section 6 concludes the paper.

2 Related Work

The task of TS is similar in nature to other sequence-to-sequence NLP tasks such as machine translation, paraphrase generation [21, 17] and most to text summarization. It can be considered as text summarization which can involve selecting sentences from the input text (extractive) or re-writing the input text (abstractive) in order to preserve most of the meaning [7]. In contrast to text summarization, simplification methods do not necessarily "compress" the input text and thus can produce longer texts, e.g. when generating term explanations. Whereas text summarization predominantly aims at filtering out the redundant text segments, TS approaches preserve the structure of the text. Despite this, a number of studies have explored the combinations of the approaches by integrating TS methods into summarization systems [27, 19].

The survey [6] provides a comprehensive overview of TS approaches, including a brief description of the earlier attempts to solve the task, discussion of various aspects of simplification (lexical, semantic, and syntactic), and the latest techniques being utilized in the field. Recent research in the field has clearly shifted towards utilizing deep learning techniques to perform TS, with a specific focus on developing solutions to combat the lack of data available for simplification. [18] is another review of the most significant studies in TS. It highlights more than 300 studies of the last three decades in the field of TS. The paper covers the corpora and evaluation metrics, for example, BLEU [13] and the most reliable metric for the sentence simplification task SARI [25].

The state-of-the-art results on TS task for English on a Turk Corpus are demonstrated by the following models:

1. DMASS & DCSS [29] is a combination of Deep Memory Augmented Sentence Simplification (DMASS) model and Deep Critic Sentence Simplification (DCSS) that has achieved 40.45 SARI.
2. ACCESS [10] by Facebook has obtained 72.54 BLEU and 41.87 SARI. The method shows that explicitly conditioning the sequence-to-sequence models on control tokens such as length, amount of paraphrasing, lexical complexity and syntactic complexity, increases the results of generation.
3. MUSS [11] has received the highest scores 78.17 (BLEU) and 42.53 (SARI). The method incorporates leveraging unsupervised data to train TS systems in multiple languages using the controllable

[1] https://newsela.com/data
[2] https://github.com/sberbank-ai/ru-gpts

generation mechanisms and pre-training.

Another line of research is focused on approaches based on reinforcement learning [28]. Transformer-based language models [23] have been applied to the sequence-to-sequence tasks for Russian, ranging from text summarization [9] to news generation[4]. The large scale pre-trained transformers represent a promising direction in the field of TS and comparable to the state-of-the-art methods. GPT-3 has achieved competitive performance on text summarization and simplification tasks [12, 22, 20]. In line with these works, we focus on the applicability of the autoregressive models, namely RuGPT3, for TS.

3 Data

The TS datasets contain parallel pairs of complex sentences (source) and their corresponding simplified versions (source).

The organizers of the RuSimpleSentEval-2021 shared task have introduced a TS dataset constructed by automatic translation and post-processed WikiLarge corpus [25]. The resulting dataset was split into train, dev and test sets. The additional dev, public and private test sets were created via crowd-sourcing using Yandex.Toloka[3]. The training set contains inappropriate examples due to being automatically constructed. Consider an example, where the sentences are likely to refer to the same town but the target sentence contains extra information which can not be derived from the source sentence: Город также является центром производства сахара и промышленности. => В 2002 году общая численность населения муниципалитета составляла 77 698 человек: 38 093 мужчины и 39 605 женщин. There are also some cases where the translation is only partially done: Belleview находится по адресу. ==> Бельвью - город во Флориде в США. Another problem is sentences where the target sentence contains more information, which is a crucial case because it contradicts the definition of simplification. The sentence is not simplified, instead it is complicated: Некоторые могут проявлять миксотроию. ==> Некоторые могут проявлять миксотрофию при использовании смешанных источников энергии. As we see further the data for training is a primary issue for the prominent performance of the TS methods. Thus, we make an attempt to overcome these issues and conduct the experiments in the following data settings: 1) all the data provided by the organizers (further in the text “$data\_all$”) 2) all cleaned data (“$clean\_all$”) 3) a 10000 examples subset of cleaned data (‘’$clean\_subset$”). The cleaning procedure of proposed data contains the following filtration steps:

- Discarding examples with less than two lemmas in the intersection between the lemmatized source and target sentences. We removed the stopwords during this step and lemmatize the sentences with pymorphy2 tagger[4];
- Discarding examples where the source sentence is a substring of the target one and the length is greater than of the source one.

4 Experimental Setup

The shared task is evaluated with SARI (System output Against References and against the Input sentence) released in EASSE[1][5]. The baseline of the competition is a multilingual BART (mBART) [8] which is commonly used for the summarization task including the Russian language [5]. The model was fine-tuned on the train set and achieved the 30.15 SARI score on the public leaderboard. We now describe the experiments conducted in this work.

Downstream transfer using pre-trained paraphrasers The motivation behind this setting is that the TS task is similar to paraphrase generation. To this end, we use the pre-trained paraphrasers for Russian and evaluate them on the task without fine-tuning [3][6]. We used mt5-base and RuGPT3 paraphrasers and the following generation hyperparameters: temperature 1, top_k repetition_penalty 1, top_p 0.9, max length 100 and the probability threshold of 0.8.

[3]https://toloka.yandex.ru/
[4]https://github.com/kmike/pymorphy2
[5]https://github.com/feralvam/easse
[6]https://github.com/RussianNLP/russian_paraphrasers

Fine-tuning Another approach includes fine-tuning of the following models:

1. mT5[26] - Multilingual T5 (mT5) by Google is a massively multilingual pre-trained text-to-text transformer model trained on the mC4 corpus in 101 languages including Russian.
2. RuGPT3-Large is a Russian open source analogue of GPT-3[2]. RuGPT3-Large[7] (760 millions of parameters) was trained on Internet text on 1024 context length with transformers on 80 billion tokens around 3 epochs, and then was fine-tuned on 2048 context.
3. RuGPT3-XL[8] was trained with 512 sequence length using Deepspeed and Megatron code by Sber-Devices team, on 80B tokens dataset for 4 epochs. After that the model was finetuned 1 epoch with sequence length 2048.

Since the best performance on the public leaderboard was achieved with the RuGPT3-XL model, we used it in a series of further experiments.

Exploring the effect of the training data size on the downstream performance In this setting, we first experiment with different sizes of the training data and the filtration procedure described in Section 3. Second, we explore the following setup:

1. **Few-shot method** with a pre-trained RuGPT3-XL model. We feed the model with 5 examples from the dev set combined with "prompts" and generate the output for the test examples. An example of the "prompt" is presented in Figure 1. For each test example, we generate 5 candidates and rank them by the lowest perplexity score.
2. **Fine-tuning and decoding methods**: we fine-tune the RuGPT3-XL model and experiment with greedy decoding, top-k and top-p sampling, and beam search methods.
3. **Post-processing** of the generated output using heuristic-based approach and re-writing the output with a pre-trained paraphraser. First, we check the appropriateness of the punctuation marks and casing of the named entities. Second, we consider the generated output to be inappropriate if: (a) the length of the output is too short, (b) there is no lemmas in the intersection of the source and target sentences, (c) the source sentence is a sub-string of the generated sentence, or the Levenshtein distance between the sentences is less than 5: in this case we rewrite the output with using the paraphraser.

Упрости: "Агрессия, как со стороны пациентов так и направленная против них, обычно случается в контексте сложных социальных взаимодействий в семье, а также является проблемой в условиях клиники и по месту жительства больного." ==> Конфликты в медучреждении - следствии конфликтов в семье.
Упрости: "Алма-Ата — лёгкая и пищевая промышленность, машиностроение; и Тараз — машиностроение, химическая и пищевая промышленность. " ==> В Алма-Ате развита лёгкая и пищевая промышленность, машиностроение; а в Таразе - машиностроение, химическая и пищевая промышленность.
Упрости: "14 декабря 1944 года рабочий посёлок Ички был переименован в рабочий посёлок Советский, после чего поселковый совет стал называться Советским." ==> 14 декабря 1944 года рабочий посёлок Ички переименован в Советский.
Упрости: "Автор гола в ворота сборной Англии, получившего название «Гол столетия», и признанного лучшим голом в истории чемпионатов мира; в той же игре забил мяч рукой, этот случай известен как «Рука Бога»." ==>

Figure 1: The "prompt" example for the few-shot technique with the RuGPT3-XL model.

All the experiments were conducted and measured on the public leaderboard. The results are presented in Table 1. During the public competition phase, the best submissions achieved with the RuGPT3-XL model trained on 10k cleaned training examples using greedy decoding and with the RuGPT3-XL model

[7] https://huggingface.co/sberbank-ai/rugpt3large_based_on_gpt2
[8] https://github.com/sberbank-ai/ru-gpts/tree/master

trained on all the cleaned data with sampling. We assumed that the combination of the two configurations was the best option. Sentence transformers[9] [15] were used to compare the generated sentences from the two configurations and to choose the best one. The library provides the multilingual model for paraphrase identification "paraphrase-xlm-r-multilingual-v1"[14]. The source sentence embeddings were compared with generated sentence embeddings from the two configurations and the one with a higher cosine similarity was kept as the final answer. Formally, the scheme of the final submission is presented in Figure 2.

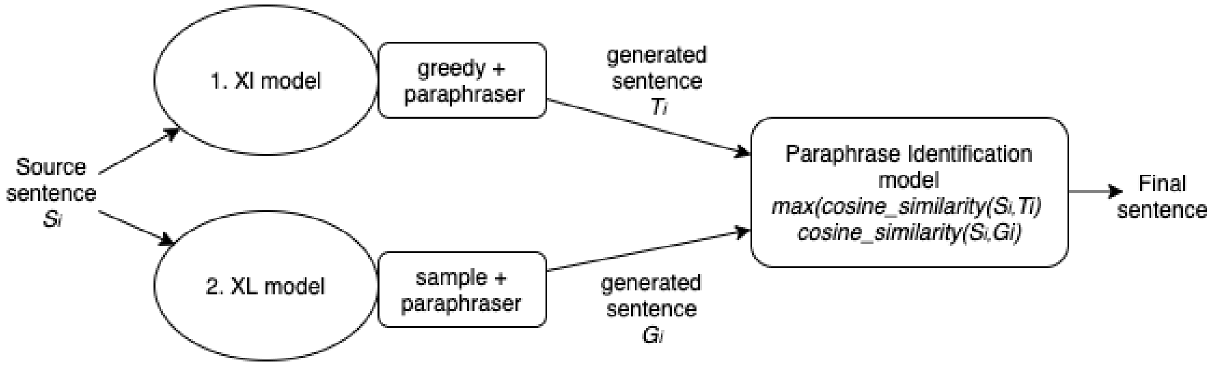

Figure 2: A graphical representation of the final method pipeline.

Method			SARI
Data	Model	Inference	
Ru/En wiki corpus	Baseline	-	30.15
Paraphraser corpus	paraphraser on RuGPT3	-	31.135
Paraphraser corpus	paraphraser on mT5	-	31.2
$data\_all$	mT5 base fine-tune	-	32.12
$clean\_all$	RuGPT3 large fine-tune	-	35.09
$nodata$	RuGPT3-XL few-shot	-	28.05
$clean\_all$	RuGPT3-XL fine-tune	beam + postproc	37.04
$clean\_subset$	RuGPT3-XL	greedy + postproc	38.10
$clean\_all$	RuGPT3-XL	sampling + postproc	38.01
$clean\_all/clean\_subset$	RuGPT3-XL	greedy/sampling + postproc	**38.87**

Table 1: Results on the public test set. **Data** represents the data on which the model was fine-tuned. **Model** shows the model, and **Inference** refers to the post-processing and decoding method.

Method			SARI
Data	Model	Inference	
$clean\_all$	RuGPT3-XL	beam + postproc	36.68
$clean\_all$	RuGPT3-XL	sample + postproc	36.78
$clean\_subset$	RuGPT3-XL	greedy + postproc	**37.82**
$clean\_all/clean\_subset$	RuGPT3-XL	greedy/sample + postproc	37.80

Table 2: Results on the private test set. **Data** represents the data on which the model was fine-tuned. **Model** shows the model, and **Inference** refers to the post-processing and decoding method.

[9]https://github.com/UKPLab/sentence-transformers

Method	Source	Generated	Estimate
few-shot	А один из ссыльных, Оскар Александрович Энгберг, изготовил жениху и невесте обручальные кольца из медного пятака.	а Один осужденный сделал молодожёнам обрученные перстни из пятикопеечной монеты	Good paraphrase, the meaning is the same
	Авиационные линии поддерживают связь крупнейших городов Италии со многими городами Европы, а также других континентов.	Крупнейшие авиарейсы позволяют экономить время для путешествий пассажиров внутри города	The thema is the same, the meaning is different.
	Актёру были близки и фарсовая, балаганная стилистика картины, и главный герой — меняющий маски и одежды подпольщик Бродский.	Актеру была близка сатирическая манера отображения происходящего его героем-подпольщиком; именно этого он сам пытался достичь за счёт изменения образа жизни (изменил имя). Неизменным оставалось лишь содержание произведения искусства - ирония!	The start of the generated sentence is good simplification, after - too many extra information.
fine-tune	В зависимости от части в сюжетной линии в прошлом времени игрок может управлять одним или двумя протагонистами.	В зависимости от части в сюжетной линии игрок может управлять одним или двумя персонажами.	Good simplification
	Аспирин — это лечение первой линии для лихорадки и симптомов суставной боли при остром ревматизме.	Аспирин обычно используется для лечения лихорадки, боли в суставах и мышечных судорог.	Good simplification
	Алкоголь может увеличить риск рождения ребёнка с врождёнными аномалиями нервной системы и обусловить задержку роста.	Алкоголь в умеренных количествах не влияет на рост и развитие плода.	Thema is the same, but the meaning is opposite
	Алма-Ата становится одним из торговых, ремесленных и сельскохозяйственных центров на Великом Шёлковом пути, имеющих монетный двор.	В Алма-Ате появляется монетный двор.	Too short simplification, the meaning is cut

Figure 3: Examples of the simplified sentences produced by the RuGPT3-XL model.

5 Results and discussion

The RuGPT3-XL models show results comparable with other methods submitted to the leaderboard. The fine-tuned models achieve results about 37 SARI depending on the inference strategy. Beam search shows worse results compared with greedy and sampling decoding methods. The performance without the post-processing component is lower across all configurations, indicating that the Russian paraphrasers are a valuable tool for simplification procedure.

We have obtained different results for the generation-based methods. First, the few-shot method is beneficial due to its simplicity. For the best result, the developer needs to investigate the prompts and choose the most optimal one. Without any fine-tuning, the RuGPT3-XL model generates a number of appropriate simplified sentences. We manually validated 50 examples produced by this method: 16% are appropriately simplified sentences, 41% are semantically inappropriate sentences but on the same topic as the original sentence, and 43% are fully inappropriate. The examples are provided in Figure 3.

The fine-tuning approach receives reasonable performance. However, there is room for improvements. One can see that the meaning of the produced sentence can be opposite despite being simplified. Another case is that the sentence gets overly "compressed" thus losing the relevant information. The best combined solution has achieved the 38.87 SARI score, as we tried to increase the score based on the best performing submissions. However this approach has not been proved to be the best option on the public leaderboard. After the competition, when the submissions were no longer limited, we discovered that the greedy decoding with post-processing shows better results. Thus, the best configuration is the RuGPT3-XL model fine-tuned on all clean_subset with greedy decoding and paraphraser post-processing that achieves a 37.82 SARI score. We observe the performance drops between the public and private test sets (from 38.10 to 37.82). A possible reason is the effect of the different generation hyperparameters for both the RuGPT3-XL model and the paraphraser, shifts in the test distributions or the model overfitting.

6 Conclusion

In this paper, we present the submission to the RuSimpleSentEval 2021 shared task devoted to the problem of text simplification. The method combines the autoregressive transformer, namely the RuGPT3-XL model, and pre-trained paraphrasers for the Russian language. The experiments are conducted using various method configurations, ranging from the few-shot and fine-tuning approaches to heuristic-based data pre-processing and post-processing procedures. The results demonstrate that the proposed method can simplify sentences with and without any fine-tuning, solely based on the prompts fed as little supervision. Our approach has achieved second place on the public leaderboard and fifth place on the private leaderboard reaching the 38.87 and 37.8 SARI score, respectively. The qualitative and quantitative analysis shows that there is still room for improvements which we consider an exciting direction for future work. Another line includes the applicability of the approach to the English language and comparison between languages. We hope that our method will be served as a prototype in the applications where text simplification is required, or used as a strong baselines for development of more sophisticated text simplification systems for Russian.

Acknowledgements

I would like to express my deepest appreciation to Vladislav Mikhailov (Sberbank, SberDevices), the best human-simplificator I know. No GPT can handle simplification tasks better than you ;).

References

[1] Fernando Alva-Manchego et al. "EASSE: Easier automatic sentence simplification evaluation". In: *arXiv preprint arXiv:1908.04567* (2019).

[2] Tom B Brown et al. "Language models are few-shot learners". In: *arXiv preprint arXiv:2005.14165* (2020).

[3] Alena Fenogenova. "Russian Paraphrasers: Paraphrase with Transformers". In: *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*. 2021, pp. 11–19.

[4] Daniil Gavrilov, Pavel Kalaidin and Valentin Malykh. "Self-attentive model for headline generation". In: *European Conference on Information Retrieval*. Springer. 2019, pp. 87–93.

[5] Ilya Gusev. "Dataset for Automatic Summarization of Russian". In: *arXiv preprint arXiv:2006.11063* (2020).

[6] Behrooz Janfada and Behrouz Minaei-Bidgoli. "A Review of the Most Important Studies on Automated Text Simplification Evaluation Metrics". In: *2020 6th International Conference on Web Research (ICWR)*. IEEE. 2020, pp. 271–278.

[7] Chandra Khatri, Gyanit Singh and Nish Parikh. "Abstractive and extractive text summarization using document context vector and recurrent neural networks". In: *arXiv preprint arXiv:1807.08000* (2018).

[8] Yinhan Liu et al. "Multilingual denoising pre-training for neural machine translation". In: *Transactions of the Association for Computational Linguistics* 8 (2020), pp. 726–742.

[9] Valentin Malykh, Denis Porplenko and Elena Tutubalina. "Generating Sport Summaries: A Case Study for Russian". In: *Analysis of Images, Social Networks and Texts*. Ed. by Wil M. P. van der Aalst et al. Cham: Springer International Publishing, 2021, pp. 149–161. ISBN: 978-3-030-72610-2.

[10] Louis Martin et al. "Controllable sentence simplification". In: *arXiv preprint arXiv:1910.02677* (2019).

[11] Louis Martin et al. "Multilingual unsupervised sentence simplification". In: *arXiv preprint arXiv:2005.00352* (2020).

[12] Takumi Maruyama and Kazuhide Yamamoto. "Extremely Low Resource Text simplification with Pre-trained Transformer Language Model". In: *2019 International Conference on Asian Language Processing (IALP)*. IEEE. 2019, pp. 53–58.

[13] Kishore Papineni et al. "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.

[14] Nils Reimers and Iryna Gurevych. "Making monolingual sentence embeddings multilingual using knowledge distillation". In: *arXiv preprint arXiv:2004.09813* (2020).

[15] Nils Reimers and Iryna Gurevych. "Sentence-bert: Sentence embeddings using siamese bert-networks". In: *arXiv preprint arXiv:1908.10084* (2019).

[16] Andrey Sakhovskiy et al. "RuSimpleSentEval-2021 Shared Task: Evaluating Sentence Simplification for Russian". In: *Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference "Dialogue"*. Vol. XX. 2021, pp. xx–xx.

[17] Advaith Siddharthan. "A survey of research on text simplification". In: *ITL-International Journal of Applied Linguistics* 165.2 (2014), pp. 259–298.

[18] Punardeep Sikka et al. "A Survey on Text Simplification". In: *arXiv preprint arXiv:2008.08612* (2020).

[19] Sara Botelho Silveira and António Branco. "Combining a double clustering approach with sen-tence simplification to produce highly informative multi-document summaries". In: *2012 IEEE 13th International Conference on Information Reuse & Integration (IRI)*. IEEE. 2012, pp. 482–489.

[20] Neha Srikanth and Junyi Jessy Li. "Elaborative Simplification: Content Addition and Explanation Generation in Text Simplification". In: *arXiv preprint arXiv:2010.10035* (2020).

[21] Suha S Al-Thanyyan and Aqil M Azmi. "Automated text simplification: A survey". In: *ACM Computing Surveys (CSUR)* 54.2 (2021), pp. 1–36.

[22] Hoang Van, David Kauchak and Gondy Leroy. "AutoMeTS: The Autocomplete for Medical Text Simplification". In: *arXiv preprint arXiv:2010.10573* (2020).

[23] Ashish Vaswani et al. "Attention is all you need". In: *arXiv preprint arXiv:1706.03762* (2017).

[24] Wei Xu, Chris Callison-Burch and Courtney Napoles. "Problems in current text simplification research: New data can help". In: *Transactions of the Association for Computational Linguistics* 3 (2015), pp. 283–297.

[25] Wei Xu et al. "Optimizing statistical machine translation for text simplification". In: *Transactions of the Association for Computational Linguistics* 4 (2016), pp. 401–415.

[26] Linting Xue et al. "mT5: A massively multilingual pre-trained text-to-text transformer". In: *arXiv preprint arXiv:2010.11934* (2020).

[27] David Zajic et al. "Multi-candidate reduction: Sentence compression as a tool for document summarization tasks". In: *Information Processing & Management* 43.6 (2007), pp. 1549–1570.

[28] Xingxing Zhang and Mirella Lapata. "Sentence simplification with deep reinforcement learning". In: *arXiv preprint arXiv:1703.10931* (2017).

[29] Sanqiang Zhao et al. "Integrating transformer and paraphrase rules for sentence simplification". In: *arXiv preprint arXiv:1810.11193* (2018).