Transfer Learning for Russian Legal Text Simplification

Mark Athugodage, Olga Mitrofanova, Vadim Gudkov
HSE University, Saint-Petersburg State University, Saint-Petersburg State University
mathugodage@hse.ru, o.mitrofanova@spbu.ru, vvagudkov@gmail.com

Abstract

We present novel results in legal text simplification for Russian. We introduce the first dataset for such a task in Russian - a parallel corpus based on the data extracted from "Rossiyskaya Gazeta Legal Papers". In this study we discuss three approaches for text simplification which involve T5 and GPT model architectures. We evaluate the proposed models on a set of metrics: ROUGE, SARI and BERTScore. We also analysed the models' results on such readability indices as Flesch-Kinkaid Grade Level and Gunning Fog Index. And, finally, we performed human evaluation of simplified texts generated by T5 and GPT models; expertise was carried out by native speakers of Russian and Russian lawyers. In this research we compared T5 and GPT architectures for text simplification task and found out that GPT handles better when it is fine-tuned on dataset of coped texts. Our research makes a big step in improving Russian legal text readability and accessibility for common people.

Keywords: Text Simplification, Text Readability, Legal text, Russian, New corpus, T5, GPT

1. Introduction

Legal documents in almost all languages are considered to be long, complex and difficult to read for people without a domain specific expertise. The texts of laws, regulations, and various resolutions are written in a very specific formal style. Legal language implies abundance of professional terms, latinisms, references to other legal documents, at the same time, these texts are considered as unemotional and syntactically complicated. It is not uncommon when a single sentence in a legal document can be a page-long (Ramaswamy et al., 2023).

The complexity of legal documents, and especially laws, complicates the life of citizens without a domain specific expertise since there is the famous Latin maxima "Ignorantia legis non excusat" ("ignorance of the law does not excuse anyone"). The only choice a simple man has is to appeal to the lawyer who may elucidate a certain law or a group of laws, but not a complete set of laws in the country.

It's notable that the government acknowledges the existence of a problem dealing with the clarity of legal documents. We can mention that the Russian Parliament recommended lawmakers use simple sentences with "SVO" structure: Subject + Verb + Object. However, some evidence suggest that Russian court resolution complexity gets even higher and higher each year (Dmitrieva, 2017).

Another significant aspect of the text complexity issue is of sociolinguistic nature: we cannot make a legal text so simple that it would be comprehensible for all citizens. The first reason concerns disabled people not all of whom are able to read and properly understand legislative documents. Then, the second reason is that the Russian Federation is a multiethnic and multilingual country, and this admits that among the citizens there may be those who do not speak Russian perfectly. Thus, the relevance of the study is explained by the high needs of society in tools and techniques for simplifying legal documents.

Since this paper is devoted to Text Simplification, it is useful to say few words about this task. Text Simplification is a text-to-text generation task, likewise summarization, machine translation, paraphrasing and style transfer. This task is often confused with summarization because of similar nature. While text summarization is always considered to be an operation of text compression, text summarization can either "compress" a text, leave it as it was or even make it larger (Fenogenova and Sberbank, 2021).

The main goal of Text Simplification is to make it easier for reader to understand a text. It becomes necessary when people without a domain specific expertise try to learn a narrow-field text, for example, a medical text. Text Simplification aims to make a specific-domain text more clear for a broader audience (Van et al., 2020) .

In the given paper we present results of research aimed at the substantiation of the possibility formal simplification of legal documents based on neural network models. We focus our attention at the development of specialised parallel legal corpus which includes the data extracted from "Rossiyskaya Gazeta Legal Papers", fine-tuning of neural models from T5 and GPT families for the simplification of legal texts and evaluation for assessing the quality of simplification. The structure of the paper is as follows: in section.

3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI), pages 59–69
20-25 May, 2024May 20th, 2024.

2. Related work

2.1. Approaches to text simplification

Modern state-of-the-art approaches for text simplification include neural-based and rule-based. Text Simplification can be performed at the lexical level, syntactic level and by means of hybrid approaches. Lexical and syntactic simplification procedures should be considered as time-tested and in most cases imply using rule-based methods. The hybrid method of text simplification is the most recent and popular at present. Yet another alternative is provided by the back-translation method. We can distinguish it as a separate class of solutions since it may be rule-based or neural-based. Although many researchers refer to backed-translation as a text simplification method, it has more in common with paraphrasing. For Russian (Galeev et al., 2021) tried back-translation as solution for a complex text: they fine-tuned a BART model for machine-translation task and then compiled "double translation". Recent works on text simplification are focused on adaptation and fine-tuning of existing neural networks - mostly Transformers. Transformers nowadays have proved to be a very efficient model for a vast list of NLP tasks - text simplification is not an exception. LSBert(Garimella et al., 2022), a Transformer-based lexical simplification model, is a bright example of lexical simplification method. LSBert finds complex words and generates the substitutions, taking into account the context. LSBert should be considered as a facilitated approach since it omits certain NLP procedures, e.g. morphological transformation. Beyond the most famous text-to-text simplification based on Transformers, there is also an edit-based method. A good example of edit-based model is EditNTS, where for each token or n-gram there are four actions offered: ADD (add token) KEEP (do not change the token; leave it as it is) DELETE (delete token) STOP If, for example, there is a sentence “She gazed at me”, then EditNTS would simplify it to “She watched me”. To make such simplification, EditNTS would need the following actions: KEEP for “she”, DELETE for “gazed”, DELETE for “at”, ADD for “watched”, KEEP for “me”, and STOP (Dong et al., 2019). There are some similar models: TST (an adaptation of GECToR corrector) (Omelianchuk et al., 2021), FELIX](Mallinson et al., 2020) and LaserTagger (Malmi et al., 2019). Such models reproduce the idea of text editing, but focus not on grammatical and orthographic errors but on simplification of complex words and phrases.

2.2. Hybrid methods using Transformers

In general there are two approaches for hybrid text simplification using neural networks: sentence-level simplification (sentence by sentence) and document-level simplification (a whole document at once). Nowadays, the most popular approach for text simplification is Transformer-based model. Transformer architecture is based on self-attention. Self-attention (also known as intra-attention) is a mechanism relating different positions of a single sequence of tokens, which makes possible the computing a representation of the sequence and modelling global dependencies. There are 3 main types of types of Transformers:

- Encoder Transformers: BERT (Devlin et al., 2018), Longformer (Beltagy et al., 2020), XL-Net (Yang et al., 2019), Transformer-XL (Dai et al., 2019);
- Decoder Transformers: GPT (Radford et al., 2019), CTRL (Keskar et al., 2019);
- Encoder-Decoder Transformers: T5 (Raffel et al., 2020), BART(Lewis et al., 2019), LED (Beltagy et al., 2020), PEGASUS (Zhang et al., 2020).

The most relevant choice for text-to-text task is encoder-decoder Transformers. We choose T5 Transformers since it is a classical example of encoder-decoder. In future, it would be reasonable to try BART as well, but BART is similar to BERT in the encoder part, which implies language masking, but at this moment BART fine-tuning with or without masking is not included in the experimental design. Then, other options for text2text generation are generative models, i.e. decoder Transformers. We use the most renowned of them - GPT, since the other model, CTRL, isn’t available for Russian yet. GPT-2 has achieved competitive performance on text summarization and simplification tasks. GPT-3 and GPT-4, as well as their modifications, are not open-source models, thus they are not available for fine-tuning. Most researchers use GPT-2 and their modifications for fine-tuning (for example, researchers used GPT-2 to fine-tune Indonesian summarizer (Khasanah and Hayaty, 2023)). Beyond GPT, there are also LLaMa and LLaMA 2, open-source LLMs from Meta AI - researchers often fine-tune these models for their specific tasks, including text simplification (Baez and Saggion, 2023).

With the emergence of large language models, NLP researchers and engineers started using prompt-engineering for many seq2seq tasks. So do they for text simplification task. People extensively use GPT-4 (Wu and Huang, 2023) with other LLMs being less popular. Although some researchers claim that transfer learning is "dead" (Pu et al.,

2023), experiments show that smaller models like BART are still perform not worse than LLMs (Sun et al., 2023). Evidence suggests that though many companies apply LLMs for their seq2seq tasks (including text simplification), smaller models are still in need, since there are some cases when one cannot train and deploy large models (Sharir et al., 2020; Chahal et al., 2020; Ahmed et al., 2023) .

3. Experimental dataset

3.1. Rationale for data selection

The crucial problem in fine-tuning seq2seq model is data availability. This problem is much more fatal for text simplification task, since there are no large datasets for this task - one can compare with a similar task, text summarization, for which there are dozen of datasets: XLSum, Newsela, CNN/DailyMail, etc. Some recent solutions are data annotation with LLMs (Gray et al., 2024). However, we find this method too risky for such a delicate field as law. Although modern LLMs are almost impeccable in performance, there is still place model hallucination as well as factual errors (Xu et al., 2024).

Since there was no dataset for Russian legal texts, we developed our own one[1]. We present dataset "Rossiyskaya Gazeta Legal Papers"[2] , which we made available on Kaggle. The dataset is based on legal papers and their simplified versions from "Rossiyskaya Gazeta" web newspaper. "Rossiyskaya Gazeta" is an official newspaper published by the Government of Russia. It's one of the widely available sources of legal documents for the citizens of Russia - the other one is a state-owned website pravo.gov.ru. Every important legal document (decisions of the High Court of Russia, Constitutional Court of Russia, orders of the President of Russia and the Government of Russia and federal laws) are published by these two sources.

In course of corpus development we selected documents accompanied by commentaries (i.e., a simplified version). The newspaper provides such commentary to what it sees as the most vital of public documents. These commentaries have legal status since they are provided by official publisher. They are aimed to serve as a simpler description for the legal document for people without a domain-specific expertise. In total our corpus has 2963 pairs of original documents and simplified ones.

[1]We used the following code https://github.com/Athugodage/RuLawSimplification/tree/main/dataset%20creation%20code

[2]https://www.kaggle.com/datasets/athugodage/russian-legal-text-parallel-corpus

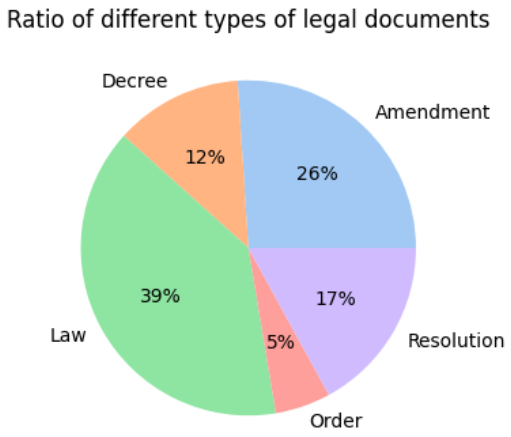

Figure 1: Types of documents in the dataset

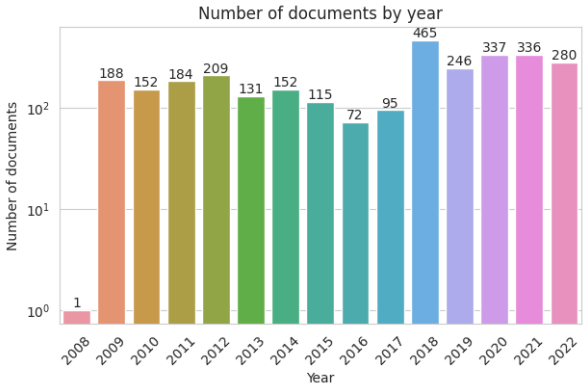

Figure 2: Distribution of the documents by year for the period from 2008 to 2022.

These documents are dated from December 2008 to November 2022. The distribution of the documents over years is shown at Figure 2.

3.2. Dataset filtration

Figure 4 clearly shows the difference in the amount of the legal documents and their simplified versions: the former are much larger than the latter. That proves the idea that the simplified version shouldn't be larger than the original text (with some exceptions), in this respect simplification is close to summarization.

When compiling the corpus, we encountered the problem of uneven distribution of documents by length, see Fig. 5 and Fig 6. E.g., the largest document of 2016 is over 100K tokens in size. In 2010, 2014, 2015 and 2019 there are documents of about 80k in size. These emissions are poorly consistent with the fact that the mean size of legal documents is about 1...2K tokens throughout the whole period. To make the dataset balanced as regards original text size - simplified text size ratio we manually fil-

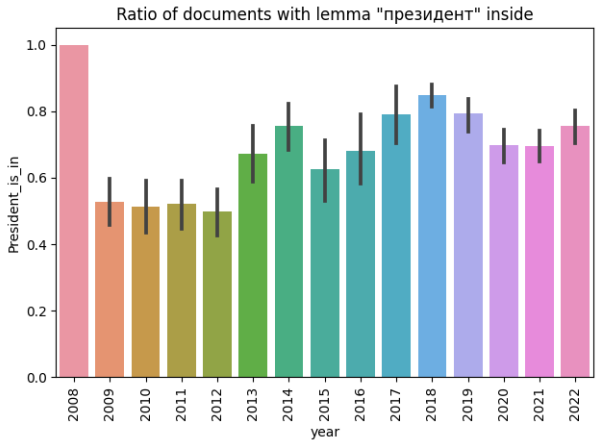

Figure 3: Ratio of documents by year where Russian lemma "president" appears

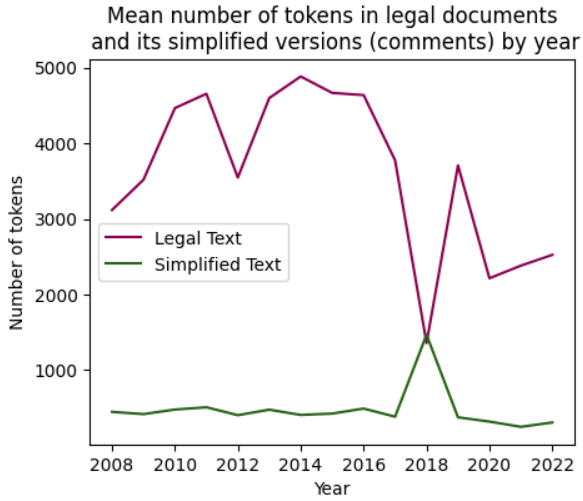

Figure 4: Mean number of tokens in legal documents set and in the set corresponding to its simplified versions (comments) by year

tered out documents of more than 40k tokens in size. We also deleted pairs of documents with the original text less than 400 tokens in size and the original text is smaller than its simplified version as the given data may lack linguistic features relevant for simplification procedure. We also believed that the original document shouldn't be smaller than its simplified version, so we deleted all pairs that fit this condition. In its final version the corpus prepared for fine tuning includes 2017 document pairs

3.3. Dataset pre-processing

For T5 model series we performed text alignment using the Natural Language Inference (NLI) model, based on RuBERT (Kuratov and Arkhipov, 2019). NLI allows us to see logical similarities between two texts. The standard model implies three-way

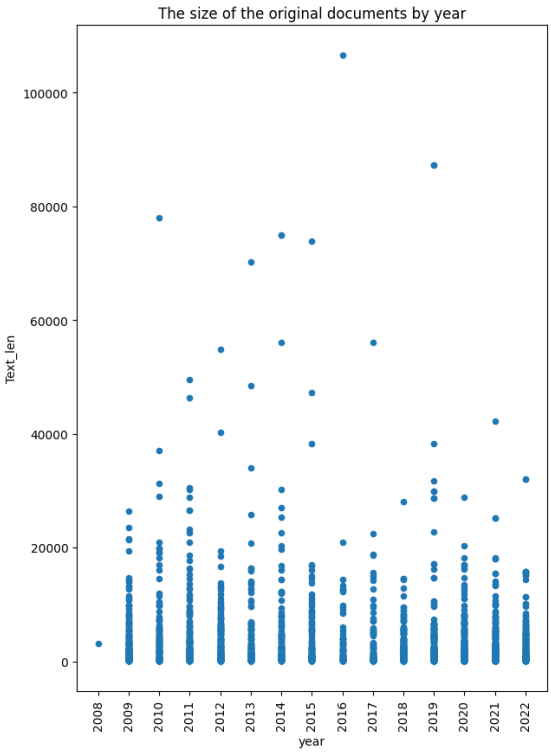

Figure 5: The distribution of document size (in number of tokens) by year

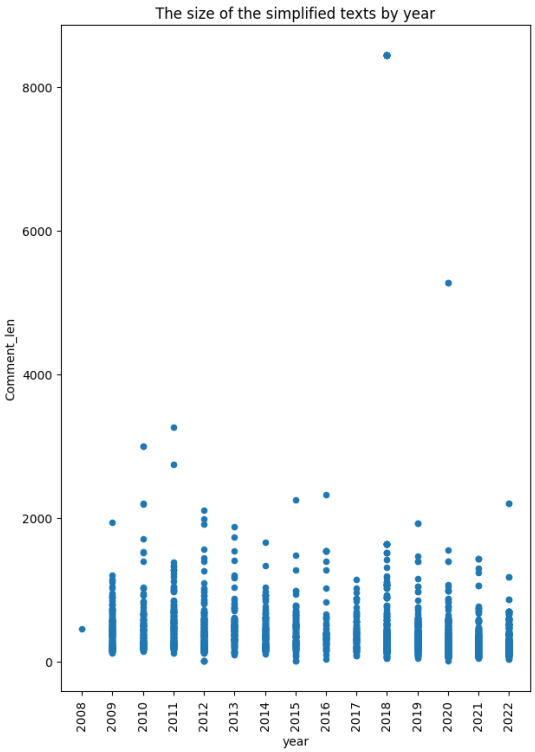

Figure 6: The distribution of simplified text size (in number of tokens) by year

inference[3]; it gives three probabilities (with values from 0 to 1): entailment (the fact that sentence B is a consequence of sentence A), neutral, contradiction (sentence B is a negation of sentence A). We used a two-way model (without neutral class) (Lin and Su, 2021). First, NLI checked that text A is a consequence of sentence B, and then vice versa. As a result, there were two entailment values - we summed them. In our case, values from 0.001 to 1.265 were obtained; sentences with mutual values less than 0.005 were determined to be slightly similar and deleted. Apart from the above described pre-processing we have also performed a custom pre-processing for GPT models. The process is described in sections below.

4. Experimental design: model selection and fine-tuning

Current T5 and GPT models for Russian do not fit text simplification task. T5 models for Russian can summarise, translate and paraphrase, but cannot simplify. Most GPT models for Russian (as well as for any other language) are intended for tasks like text generation, question answering and chatting, though some researchers tried to teach GPT simplify in Russian (Shatilov and Rey, 2021). The newest Open AI's ChatGPT-4, Yandex's YaGPT and Sber's Gigachat can simplify a text if a user asks it (however, there are still considerations on the quality of such simplification). This is why we decided to fine-tune our own models. In the following sections you can read about our fine-tuned models: *T5-RLS2000, GPT-simplifier-large-text, and GPT-simplifier25*. They are based on mainstream Russian models from Sber.

4.1. Larger T5 model (*T5-RLS2000*)

This model [4] is based on the Russian-language model T5 from Sber (Zmitrovich et al., 2023) on the entire aligned body of 2 thousand pairs of articles. The fine-tuning was conducted at a rate of 0.00002 on 3 epochs. More information is available in the model's card. This model cannot process multi-sentence texts - one may enter just one sentence in the input.

4.2. Larger GPT model (*GPT-simplifier-large-text*)

This model [5] is based on the Russian GPT3 model from Sber, which in turn is a trained GPT-2 model from OpenAI. The model is fine-tuned on a standard case of 2 thousand pairs of articles. The texts were submitted in full form without compression. The model is trained on 10 epochs with a learning rate of 0.00005. For faster and more efficient operation, gradient accumulation was used every 8 moves.

4.3. Smaller GPT model (*GPT-simplifier25*)

Working with the above mentioned models, we came to the conclusion that the main problem of legal text processing and simplification is the large size of the documents. This problem could be solved if we filter the document. The first sentence in every legal document is introductory *(Examples: "*Именем Российской Федерации*" -> "In the name of Russian Federation"; "*Принят Государственной Думой*" -> "Adopted by State Duma")*. The second phrase corresponds to the pattern like the following: *"*Конституционный Суд Российской Федерации в составе Председателя Х, судей А, Б, В, Г, Д, Ж, руководствуясь статьей 100 Конституции Российской Федерации, пунктом 1 статьи 2 Гражданского Кодекса Российской Федерации [...]*" (in English: "Russian Constitutional Court, consisting of the Chairman X, judges A, B, C, D, E F, [made a decision] in accordance of article 100 of the Russian Constitution, paragraph 1 of the article 2 of the Russian Civil Code, [and so on... This listing can be page-long]")*. We skip these two sentences.
Also, with the help of regular expressions, sentences with too long references to other laws were removed. For example, it is common in Russian legal texts to give citation in brackets just in the middle of the sentence like this: *"*"(постановления от 30 октября 2003 года N 15-П, от 27 июня 2012 года N 15-П, от 18 июля 2013 года N 19-П и др.)*"*. We delete it.
This allowed us to examine a clear text without citation and unnecessary phrases. If the document was still too big (e.g. the document had more than 40 sentences), we left just last 35 sentences (removing all others). This action may seem controversial for some researchers, since one can claim that we let significant context be left aside. But that is not true, since the structure of Russian legal document itself is designed so that the most informative part is **always** left in the end of the document. The beginning of any document has somewhat a ritualistic nature. It is almost always filled

[3] https://huggingface.co/cointegrated/rubert-base-cased-nli-threeway

[4] https://huggingface.co/marcus2000/T5-RLS2000

[5] https://huggingface.co/marcus2000/GPT_simplifier_large_text

with some phrases like those mentioned above. In contrast, all important decisions, terms and regulations traditionally placed in the end. Thus, cutting text leaving just last 35 sentences did not affect completeness of the content. After that, fine-tuning of the same model of Sber was carried out. The new model[6] was named *GPT-simplifier25*, because it was trained on 25 epochs. Comparing these two GPT models may be scientifically interesting, since it shows whether text reduction is possible (in our case) and, if it is, whether the model which was fine-tuned on dataset with reduced texts has better results than the other one. We did this to check a hypothesis that data economy could positively impact on the result. The following document [7] is a good example of our claim: the document itself starts with some external information, then the first paragraph is the argumentation of the order; the real content starts from the second paragraph.

5. Automatic evaluation

Automatic metrics for simplification include primarily SARI and SAMSA (Grabar and Saggion, 2022). In addition, there is a number of metrics that are often used to evaluate simplification, but in fact they are common for any seq2seq task in NLP. For example, ROUGE is almost always mentioned in similar studies, but this metric was originally designed for summarization (Lin, 2004). There are some other rare metrics which were primarily designed for a specific contest, as with RuSimScore, which was introduced during RuSimpleSentEval (RSSE) in 2021 (Orzhenovskii, 2021). Having analysed different groups of metrics, we focused our attention on ROUGE, BERTScore and SARI. To evaluate each of the pre-trained models, we proposed our own approach. GPT models were evaluated on a set of 2500-characters-long excerpts (because these models cannot have a limited context) from original documents (from the test set). T5 model were evaluated on the test set of the aligned sentences: the algorithm checked to what extent the model simplifies each sentence. We see that on ROUGE metrics our models show bad results, comparing to summarization models. The best our model in this case is GPT-large. On another equally interesting BERTScore metric, the best results are obtained for our T5 model. Still, the most prominent metric for us remains SARI, since only this (of proposed ones) shows the real efficiency of the text simplification model. The best result (54.96) on SARI belongs to *T5-RLS2000*. This is an excellent result; for comparison, in 2021 the state-of-the-art result in text simplification was 44.3(Omelianchuk et al., 2021). In some other works the SARI score is around 35 (Sun et al., 2021) . Further evaluation of the simplification abilities of the models was performed using readability indices. We examined a set of resulting simplified texts from our fine-tuned GPT models and selected Gunning Fog Index and Flesch-Kincaid Readability Index to evaluate them. We made our own script to evaluate these indices because standard versions of them that are available in open-source Python libraries, are more suitable for English. Our version of these formulas allow take into account the specificity of Russian text. Table 2 shows the results of checking simplified texts from the test sample on the Gunning Fog Index. The table shows the average number for 100 documents. The Gunning Fog Index gives a difficulty score for each text individually. The table below shows the complexity index of the original and the text simplified by a specific model.

The same table shows the results of checking the Flesch-Kinkaid readability index in the values of the training classes, i.e. how much you need to study (on average) to understand this or that text. As can be seen from the two tables with estimates of the readability indices, the small GT3 model copes with simplification much better than the large one (Blinova and Tarasov, 2022). The T5 models did not participate in the evaluations on the readability index, because these are simplification models based on proposals. However, we offer table 3 to show the readability estimates for other T5 models for summarization and paraphrasing. Such a comparison is also interesting because it clearly shows the fundamental difference between the task of simplification and summarization.

6. Human evaluation

For a more qualified evaluation we asked 20 respondents to access simplified texts generated with our models. The respondents were presented with four legal documents:

- Federal law dated 30.12.2021 № 454-FZ “About seed production”[8]
- Resolution of the Chief State Sanitary Doctor of the Russian Federation dated 02.07.2021 No. 17 "On Amendments to the Resolution of the Chief State Sanitary Doctor of the Russian Federation dated 03/18/2020 No. 7 "On ensuring the isolation regime in order to prevent the spread of COVID-2019"[9]

[6] https://huggingface.co/marcus2000/GPT_simplifier25

[7] http://publication.pravo.gov.ru/Document/View/0001202210170033

[8] http://publication.pravo.gov.ru/Document/View/0001202112300119

[9] http://publication.pravo.gov.ru/Document/View/0001202107060020

Table 1: Automatic evaluation using ROUGE, BERTScore, SARI.

Metric	ROUGE				BERTScore		SARI
Model	1	2	3	LSUM	P[a]	F1 [b]	SARI
T5-RLS2000	5	0.6	0.05	5	**0.65**	0.64	**54.96**
GPT s. 25	2.1	0	1.79	1.84	0.61	0.6	40.96
GPT s. large	7.16	1.25	6.7	6.85	0.61	0.6	39.9
rut5 base sum gazeta	9.25	**2.39**	**9.2**	**9.39**	0.6	0.6	35.5
ruT5 large	**10.2**	0.5	**9.2**	9.2	0.6	0.58	34.51
mbart ru sum gazeta	7	1.16	6.59	6.54	-	-	53.9
rut5 base paraphraser	3.3	0.22	2.47	2.44	0.53	0.53	35.62

[a] Mean Precision in BERTScore metric
[b] Mean F1 score in BERTScore metric

Table 2: Gunning Fog Readability and Flesch-Kincaid Grade Level Readability Indices evaluations on our models

	Gunning Fox Index		FKGL	
Model	Original	Simplified	Original	Simplified
GPT simplifier 25	**59.5**	**41.8**	**26.58**	**18.23**
GPT simplifier large	59.5	54	26.58	25.48

- Federal law dated 03.04.2023 N 108-FZ “About making changes to Federal Law “On State Regulation of Production and Turnover of Ethyl Alcohol, Alcoholic and Alcohol-Containing Products and on Restriction Consumption (Drinking) of Alcoholic Products”[10]
- Federal law dated 21.11.2022 № 455-FZ “On amendments to Federal Law “On State Benefits to Citizens with Children”[11]

Each of the documents had five simplified versions (four of them generated with our T5 and GPT models [12], one being the commentary from Rossiyskaya Gazeta newspaper). Respondents were asked to rate each text with a score from 0 to 10. During the evaluation, respondents were recommended to give special priority to the following criteria:

- literacy,
- readability (easy to read, no complicated lexical items)
- conveys the basic principles of the document (the more specified, the better),
- authenticity of facts.

The assessment was conducted in the form of a survey in Google Forms. The results were evaluated in two groups of respondents – in a group of experts with a degree in law (or, at least, a law student), and in a group of experts without a legal education. Eventually, 20 people took part in the survey. Among them five respondents confirmed their qualification in legal sciences, 15 respondents

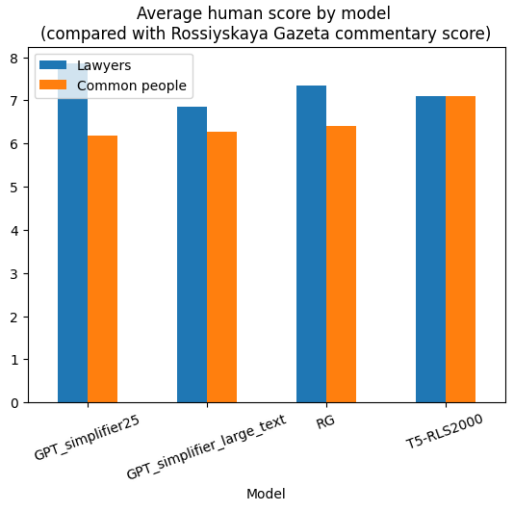

Figure 7: Average score of the human evaluation for each of the models. Orange columns (the left-side graph) represent the scores among lawyers; green columns - among people without a domain specific expertise

turned out to be non-specialists in law. On average, the assessment of legal experts is 1 point higher than experts without a degree in law. Judging from Fig. 7, it can be concluded that lawyers consider the model *GPT-simplifier25*, trained on abbreviated texts, as the most accurate. The rest consider the simplification model according to the proposals of *T5-RLS2000* to be the best. It should be noted that in both cases, the proposed models got higher ratings than the Rossiyskaya Gazeta commentary.

In general, the fact that the proposed neural network models coped with a lot of documents better than comments written by a living person can be

[10] http://publication.pravo.gov.ru/Document/View/0001202304030011
[11] http://publication.pravo.gov.ru/Document/View/0001202211210043
[12] Examples are given in Appendix A.

Table 3: Gunning Fog Readability and Flesch-Kincaid Grade Level Readability Indices evaluations on other models (for comparison)

	Gunning Fox Index		FKGL	
Model	Original	Simplified	Original	Simplified
rut5-base-sum-gazeta (summarization)	53.5	62.7	26.58	29.04
ruT5-large (summarization)	53.5	36.56	26.58	10.86
rut5-base-paraphraser (paraphrasing)	53.5	137.9	26.58	126

considered a success of experiments on fine-tuning simplification models for Russian legal texts.

7. Conclusion

In this article we discussed the problem of automatic simplification of Russian legal texts. The work presents three new fine-tuned neural network models: T5-based and GPT-based. In order to fine-tune the models we developed a new parallel corpus based on Russian legal documents and commentaries. This corpus contains a pair of an original legal text and its description, provided by Rossiyskaya Gazeta (a newspaper published by the Government of Russia). The discussed language models have significant differences since the size of models' datasets varied a lot. The models were evaluated with ROUGE, SARI and BERTScore. The generated texts were analysed as regards readability indexes Flesch-Kinkaid Grade Level and Gunning Fog Index. We asked 20 respondents to participate in human evaluation of the fine-tuned models.

The proposed solutions take a big step in expanding the availability and readability of legal documents for wide audience. With the help of the proposed models, it is possible to simplify professional legal texts so that they can be understood by almost everyone. However, at this stage, simplified texts may have some shortcomings, thus, verification of the simplified texts by experts or editors may be required. Our next challenge is to improve existing simplification technology so that the user could read generated texts immediately after the procedure. The future work deals with fine-tuning Longformer Encoder-Decoder and LongT5 for simplification task and with reduction of defects in generated texts.

8. Acknowledgments

This research was supported in part through computational resources of HPC facilities at HSE University.

9. Bibliographical References

References

Ishrat Ahmed, Yu Zhou, Nikhita Sharma, and Jordan Hosier. 2023. Text summarization for call center transcripts. In *Intelligent Systems Conference*, pages 542–551. Springer.

Anthony Baez and Horacio Saggion. 2023. Lsllama: Fine-tuned llama for lexical simplification. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 102–108.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Olga Blinova and Nikita Tarasov. 2022. A hybrid model of complexity estimation: Evidence from russian legal texts. *Frontiers in Artificial Intelligence*, 5:248.

Dheeraj Chahal, Ravi Ojha, Manju Ramesh, and Rekha Singhal. 2020. Migrating large deep learning models to serverless architecture. In *2020 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW)*, pages 111–116. IEEE.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Arina Dmitrieva. 2017. Art of legal writing: quantitive analysis of the resolutions of the constitutional court of russian federation. *Comparative Constitutional Review (Saint-Petersburg, Russia)*, 3:125–133.

Yue Dong, Zichao Li, Mehdi Rezagholizadeh, and Jackie Chi Kit Cheung. 2019. Editnts: An neural programmer-interpreter model for sentence simplification through explicit editing. *arXiv preprint arXiv:1906.08104*.

Alena Fenogenova and SberDevices Sberbank. 2021. Text simplification with autoregressive models. *Proc. Computational Linguistics and Intellectual Tech*, pages 1–8.

Farit Galeev, Marina Leushina, and Vladimir Ivanov. 2021. rubts: Russian sentence simplification using back-translation. *Proc. Computational Linguistics and Intellectual Tech*, pages 1–8.

Aparna Garimella, Abhilasha Sancheti, Vinay Aggarwal, Ananya Ganesh, Niyati Chhaya, and Nanda Kambhatla. 2022. Text simplification for legal domain:{l} nsights and challenges. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 296–304.

Natalia Grabar and Horacio Saggion. 2022. Evaluation of automatic text simplification: Where are we now, where should we go from here. In *Traitement Automatique des Langues Naturelles*, pages 453–463. ATALA.

Morgan A Gray, Jaromir Savelka, Wesley M Oliver, and Kevin D Ashley. 2024. Empirical legal analysis simplified: reducing complexity through automatic identification and evaluation of legally relevant factors. *Philosophical Transactions of the Royal Society A*, 382(2270):20230155.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Aini Nur Khasanah and Mardhiya Hayaty. 2023. Abstractive-based automatic text summarization on indonesian news using gpt-2. *JURTEKSI (Jurnal Teknologi dan Sistem Informasi)*, 10(1):9–18.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language. *arXiv preprint arXiv:1905.07213*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yi-Chung Lin and Keh-Yih Su. 2021. How fast can bert learn simple natural language inference? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 626–633.

Jonathan Mallinson, Aliaksei Severyn, Eric Malmi, and Guillermo Garrido. 2020. Felix: Flexible text editing through tagging and insertion. *arXiv preprint arXiv:2003.10687*.

Eric Malmi, Sebastian Krause, Sascha Rothe, Daniil Mirylenka, and Aliaksei Severyn. 2019. Encode, tag, realize: High-precision text editing. *arXiv preprint arXiv:1909.01187*.

Kostiantyn Omelianchuk, Vipul Raheja, and Oleksandr Skurzhanskyi. 2021. Text simplification by tagging. *arXiv preprint arXiv:2103.05070*.

Mikhail Orzhenovskii. 2021. Rusimscore: unsupervised scoring function for russian sentence simplification quality. In *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, pages 524–532.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. *arXiv preprint arXiv:2309.09558*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Sankar Ramaswamy, R Sreelekshmi, and G Veena. 2023. Complexity analysis of legal documents. In *International Conference on Artificial Intelligence on Textile and Apparel*, pages 141–154. Springer.

Or Sharir, Barak Peleg, and Yoav Shoham. 2020. The cost of training nlp models: A concise overview. *arXiv preprint arXiv:2004.08900*.

AA Shatilov and AI Rey. 2021. Sentence simplification with rugpt3. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue*, pages 1–13.

Renliang Sun, Hanqi Jin, and Xiaojun Wan. 2021. Document-level text simplification: Dataset, criteria and baseline. *arXiv preprint arXiv:2110.05071*.

Renliang Sun, Wei Xu, and Xiaojun Wan. 2023. Teaching the pre-trained model to generate simple texts for text simplification. *arXiv preprint arXiv:2305.12463*.

Hoang Van, David Kauchak, and Gondy Leroy. 2020. Automets: the autocomplete for medical text simplification. *arXiv preprint arXiv:2010.10573*.

Shih-Hung Wu and Hong-Yi Huang. 2023. A prompt engineering approach to scientific text simplification: Cyut at simpletext2023 task3.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Dmitry Zmitrovich, Alexander Abramov, Andrey Kalmykov, Maria Tikhonova, Ekaterina Taktasheva, Danil Astafurov, Mark Baushenko, Artem Snegirev, Tatiana Shavrina, Sergey Markov, Vladislav Mikhailov, and Alena Fenogenova. 2023. A family of pretrained transformer language models for russian.

10. Appendix A. Examples

Original text: Link to full text: https://rg.ru/2022/01/11/semenovodstvo-dok.html

Rossiyskaya Gazeta simplified version: Новая редакция закона о семеноводстве поможет в том числе побороть импортозависимость России от зарубежных семян. Закон об этом публикует "Российская газета". Закон о семеноводстве не касается дачников, использующих семена для собственных нужд. Но он важен для российского АПК. Документ, в частности, регламентирует основные процессы по производству, хранению, реализации, транспортировке и использованию семян растений, а также по их импорту и экспорту из России. В новой редакции определены базовые понятия ("семена "сорт "гибрид"и многие другие). Закон запрещает ввозить и использовать для посевов в России семена, содержащие ГМО. Для борьбы с фальсификатом будет создана федеральная государственная информационная система. Документ предполагает создание госреестра селекционных достижений, допущенных к использованию. Кроме того, в новой редакции предусмотрена необходимость оформления удостоверений о качестве семян сельхозрастений. Предыдущая версия закона была принята еще в 1997 году и нуждалась в обновлении. Новая редакция долго и бурно обсуждалась в отраслевом сообществе. В результате в окончательную версию были внесены принципиальные изменения - исключены необходимость лицензирования и сертификации семян как избыточные меры. Лицензирование заменяется реестром производителей семян, сертификация - удостоверением качества. Закон должен создать условия для развития рынка семян в России и повышения их качества. Власти неоднократно обращали внимание на слабую обеспеченность российского АПК отечественными семенами. Семена некоторых культур полностью завозятся из-за рубежа. Новые нормы закона вступят в силу с 1 сентября 2023 года.

GPT-simplifier-large-text: В России с 2021 года вводится новый вид разрешенного использования семян - "сорт-сорт". В России он появился в конце прошлого года. По новому закону, если в одном регионе выращивается более трех тысяч сортов и гибридов сельскохозяйственных растений, то в каждом регионе будет действовать свой вид разрешенного использования семян.

GPT-simplifier25: В России впервые появился единый госреестр сельхозкультур с правами экспонирования до 2015 г., уточнили "РГ"во время обсуждения документа минсельхозпрода РФ Дмитрием Медведевым после заседания правительственной комиссии при Росреестре 25 ноября прошлого году о развитии отрасли сельского хозяйства... Подробнее об этом можно прочесть здесь. Сегодня же этот порядок продлен еще более чем наполовину, говорит замдиректора департамента агропродовольственных отношений Минсельхоза Михаил Блинкин ("РГ"). По словам эксперта ведомства Алексея Сидорова : это связано со снижениемелностью цен производителей сельскохозяйственно й продукции вследствие снижения ее доли импорта как сырья... Это происходит потому что производители меньше закупают импортной техники у фермеров перед началом сезона хранения урожая либо перепрофилируют свои машины под хранение влаги осенью этого периода без необходимости проведения дополнительных работ даже если урожай уже созрели полностью.. При снижении себестоимости выпускаемой аграрной культуры затраты производителя могут быть уменьшены только благодаря снижению затрат посредников между производителями тракторов зернохранилищ ввиду отсутствия перекрестников обработки зерна через один агрегатный аппарат."Это значит лишь то, -что новые правила распространяются исключительно непосредственно именно сейчас!

А раньше они были введены просто так – чтобы поддержать спрос населения сказал он.

T5 RLS2000: В России вводится режим самоизоляции. В России вводится режим самоизоляции. В России вводится генетическая экспертиза семян. В России вводится новый порядок лицензирования сельскохозяйстве нных предприятий. В частности, в законе о семеноводстве. В частности, в отношении сельскохозяйстве нных культур. В России вводится система агрострахования. В частности, в области семеноводства. В России будет создана единая государственная политика в области семеноводства. В частности, в области семеноводства. В частности, в России будет создано министерство. сельского хозяйства. В России будет создана государственная информационная система в области семеноводства. В частности, в России будет создан банк стандартных образцов семян. В частности, в России будет создано министерство сельского хозяйства. В частности, в отношении семян сельскохозяйстве нных культур. В частности, в области семеноводства. В частности, в ходе проведения лабораторно - биологических исследований. В частности, за счет бюджета регионов. В частности, в области семеноводства. В документе, в частности - о порядке проведения научных исследований. В России вводится система агрострахования. В зависимости от этапа производства семян сельскохозяйстве нных культур определяются категории агрокультур