Grammatical acceptability evaluation using the material of the RuCoLA corpus

Yan Prourzin

September 2023

Abstract

The purpose of this work is to find a qualitative method for assessing the grammatical acceptability of sentences in Russian. As part of the work, the task of using marked-up corpora to train a model working with morphological features of words was solved. The work demonstrates the high potential of using marked-up linguistic corpora to solve natural language processing problems.

`https://github.com/yanpr0/NLP-course-project`.

1 Introduction

In recent years, a significant progress has been made in the field of natural language processing (NLP) due to the emergence of new neural network models that can be trained on untagged data, which makes it possible to train on truly huge amounts of information. One of the tasks is an assessment of the acceptability of sentences. In linguistics, acceptability is used to evaluate a sentence on a scale from possible to impossible from the point of view of a native speaker. The following main characteristics are introduced to indicate acceptability:

- Meaningful – semantically correct text, free from logical errors.
- Grammatical – a text that satisfies the rules of grammar of the language. This may include the correctness of the word order, the correspondence of morphological features, syntax.

A meaningful and grammatical text is acceptable.

Having a working solution to the problem of assessing acceptability, it is possible to achieve improvements in other NLP tasks: for example, at the moment there are already models capable of qualitatively solving machine translation or summarization problems, but even the most modern neural networks are far from perfect and there may be various errors and deficiencies. To combat this, you can use a solution to assess the acceptability and arrange the outputs of the model. The article [Batra et al., 2021] demonstrates that this method is able to

improve the quality of generated texts. In this work, emphasis was placed on the use of morphological characteristics to achieve better results in the classification of sentences with morphological and syntactic errors.

1.1 Team

Yan Prourzin.

2 Related Work

To solve the problem of assessing acceptability by a team of RuCoLA researchers several machine learning models of varying degrees of complexity were prepared and tested as a baseline. The problem of binary classification was solved, where the *acceptable* parameter of the RuCoLA dataset was used as a label. Among the simple and basic models were selected:

- Majority vote – a model that finds the most common label of the predicted class parameter and predicts exactly this label on any input.
- Logistic regression is a method of constructing a linear classifier that allows you to estimate probabilities of belonging of objects to classes. In this case, as a matrix of features, we use the matrix $TF\text{-}IDF$, calculated on all the sentences of the corpus.

Models based on the Transformer [Vaswani et al., 2017] were also tested – models using the attention mechanism [Bahdanau et al., 2014] to increase the learning rate: T5 [Raffel et al., 2019], XLM-R [Conneau et al., 2019], BERT [Devlin et al., 2018], RemBERT [Chung et al., 2020], RoBERTa [Liu et al., 2019]. Recent works on this problem shows that solutions based on transformers show the best results.

3 Model Description

We will pre-train a model that can work with grammatical features of words and at the same time take both the right and left context into account to capture syntactic connections in a sentence. After that, this pre-trained model can be used in training as part of the main model. For pre-training, we will use an open corpus of morphological markup of sentences with the homonymy removed. For the convenience of working with morphological features, we will use the *rnnmorph*[Gusev, 2021] library. With its help, we will bring the data into a form that *rnnmorph* can work with (see the example [1]).

We will use the features: POS – part of speech, and Grammemes – morphological characteristics. Let's use the model with the BERT architecture, but modify the tokenizer: the standard approach uses tokenization using the WordPiece algorithm, but we will do it differently: we will collect all possible combinations of grammatical features from the corpus and supplement them with other combinations possible due to homonymy (we will get them using the

Token	Norml form	POS	Grammemes
«	«	PUNCT	_
Школа	школа	NOUN	Case=Nom|Gender=Fem|Number=Sing
злословия	злословие	NOUN	Case=Gen|Gender=Neut|Number=Sing
»	»	PUNCT	_
учит	учить	VERB	Mood=Ind|Number=Sing|Person=3|...
прикусить	прикусить	VERB	VerbForm=Inf
язык	язык	NOUN	Case=Acc|Gender=Masc|Number=Sing

Table 1: Fragment of dataset after using rnnmorph

*pymorphy*2 [Korobov, 2015] library). Obtained string representations of grammatical characteristics together with the standard tokens for BERT ([PAD], [UNK], [CLS], [SEP], [MASK]) will make up the tokenizer dictionary. With the help of the tokenizer, we can encode sentences from the corpus – and then we have prepared data for pre-training. We will train our model only on one of the two tasks on which the original BERT is pre-trained – on the MLM task (Masked Language Modeling), since we need to catch dependencies only inside the sentence. In this work, BERT and tokenizer implementations from the Transformers[Wolf et al., 2020] library were used.

3.1 Model architecture

Let's choose the following architecture for our main model [1]:

- getting grammatical characteristics of an input sentence using *rnnmorph*;
- tokenization: sentences are tokenized for the selected model (ruRoBERTa-large is taken as the best of baselines), grammatical characteristics are tokenized with a tokenizer for a pre-trained model;
- concatenation of the pooler's outputs of both models and passing them through the linear layer to obtain logits.

4 Dataset

The RuCoLA corpus (Russian Corpus of Linguistic Acceptability) consists of sentences supplied with a binary acceptability label, the type of error made in the example, and the source from which the sentence was received. In total, there are 4 types of errors in the corpus: morphological, syntactic, semantic (to test models on standard phenomena) and hallucinations (to check the quality of the model on machine-generated texts) (Fig. 2).

In order to use the corpus to improve the quality of language models, two types of sentences were added: written and arranged by linguistic experts, as well as generated by language models. The sentences of the first type, together

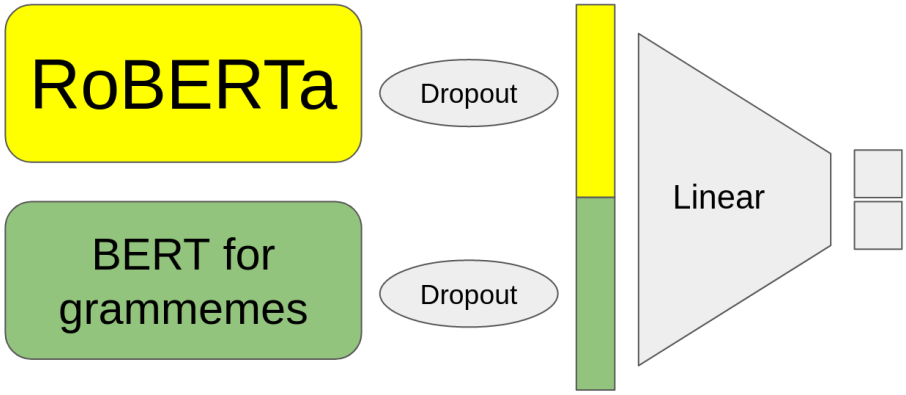

Figure 1: Model architecture

with the acceptance marks affixed by experts, were taken from the works of linguists on various aspects of the language.

Example (Yakov Testelets, "Introduction to the general syntax"):

Я обнаружил его лежащего одного на кровати – приемлемое;

Я хочу, чтобы они зайдут, когда ты спрячешься – неприемлемое.

The sentences of the second type were obtained differently: by applying paraphrase and machine translation models on various corpora (Tatoeba, TED, WikiMatrix, parallel Yandex.Translator corpus), sentences were then assigned the acceptance markup on Yandex.Toloka, subsequently validated by students of respective specialties with an indication of the specific type of errors from the above.

Example:

Хочешь, я отвезу тебя в аэропорт? (machine translation model) – acceptable;

Это всю историю была исполнения в живую телестудию (paraphraser) – unacceptable.

	sentence
0	Диктат традиций оказался очень силён.
1	Он спит мимо сада.
2	"Он мгновенно понял, что произошло что-то не то, причем очень изумился, потом повозмущался и
3	Вторая существовала с 1852 по 1973 год и была объединена с частими округа Дурхам в Дурхамски
4	"Футболист ""Ливерпуля"" впервые подписал первый год поражения колена."

Table 2: RuCoLA data example

To train a model capable of working with grammatical features, a corpus was used, obtained by bringing it to a general form and combining the corpora with manual morphological markup from OpenCorpora and General Internet-Corpus of Russian.

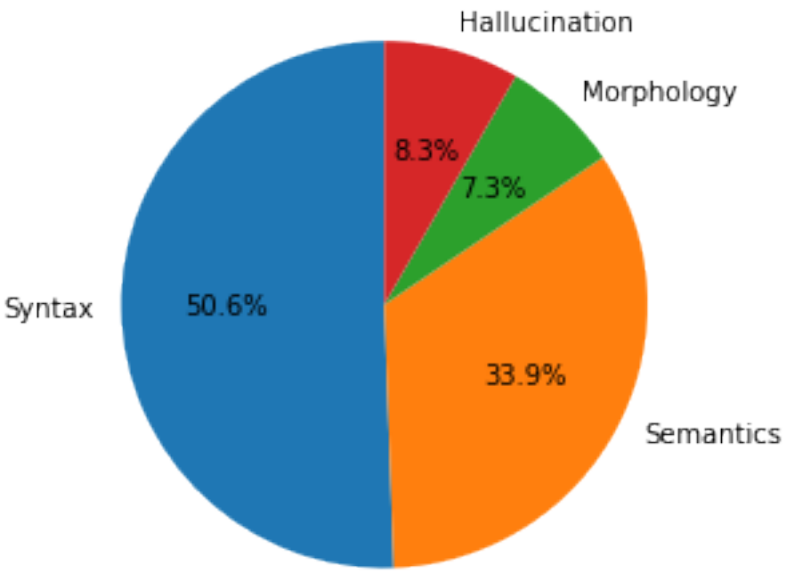

Figure 2: Types of errors in RuCoLA dataset

5 Experiments

5.1 Metrics

The following metrics were used to rank the models:

- *Accuracy* – the percentage of correctly assigned labels among all model responses.
- MCC (Matthew's correlation coefficient) – a statistical value that allows you to estimate the proximity of two binary quantities. Suppose we have a standard confusion matrix (Fig. 3), then

$$MCC = \frac{TN \times TP - FN \times FP}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

.

5.2 Experiment Setup

The pre-training of the model working with grammemes was conducted using the following hyper-parameters:

- train/validation data size = 9/1

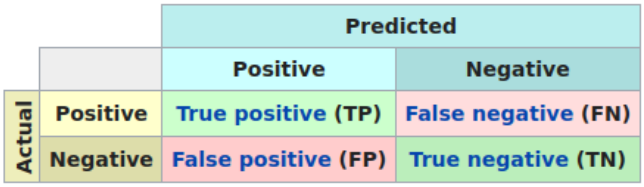

Figure 3: Confusion matrix

- number of train epochs = 5,
- train batch size = 32,
- learning rate = $1e-5$,
- weight decay = 0.0001,

During the training of the main model different hyper-parameters were tested:

- learning rate: $e-5$, $3e-5$, $5e-5$,
- weight decay: $1e-4$, $1e-2$, 0.1,
- batch size: 32, 64.

Turned out the optimal number of epochs is 5, after which the model begins to overtrain. With a smaller batch size, the model learns noticeably better than with a larger one (32 and 64 in the experiments conducted). There were 3 runs with different seed for each set of hyper-parameters.

5.3 Baselines

Rank	Team	Model	Acc	MCC
1	RuCoLA Team	ruRoBERTa-large	0.80	0.53
2	RuCoLA Team	ruBERT-large	0.77	0.46
3	RuCoLA Team	RemBERT	0.76	0.43
4	RuCoLA Team	ruBERT-base	0.75	0.41
5	RuCoLA Team	ruT5-large	0.70	0.28
6	RuCoLA Team	XLM-R-base	0.72	0.28
7	RuCoLA Team	ruT5-base	0.70	0.23
8	RuCoLA Team	Logistic Regression + TF-IDF	0.67	0.04
9	RuCoLA Team	Majority Vote	0.68	0

Table 3: Baseline metrics from RuCoLA team: aggregate rating

We see that the ruRoBERTa-large model showed itself best in both cases.

Rank	Team	Model	Expert		Machine	
			Acc	MCC	Acc	MCC
	RuCoLA Team	Human Benchmark	0.84	0.57	TBA	TBA
1	RuCoLA Team	ruRoBERTa-large	0.83	0.51	0.79	0.53
2	RuCoLA Team	ruBERT-large	0.80	0.43	0.76	0.46
3	RuCoLA Team	RemBERT	0.79	0.42	0.74	0.43
4	RuCoLA Team	ruBERT-base	0.79	0.40	0.74	0.41
5	RuCoLA Team	ruT5-large	0.75	0.35	0.68	0.25
6	RuCoLA Team	XLM-R-base	0.77	0.29	0.69	0.26
7	RuCoLA Team	ruT5-base	0.78	0.32	0.67	0.18
8	RuCoLA Team	LogReg + TF-IDF	0.76	0.17	0.63	-0.02
9	RuCoLA Team	Majority Vote	0.74	0	0.65	0

Table 4: Baseline metrics from the RuCoLA team: separate ratings for expert and machine-generated sentences

6 Results

This approach has managed to achieve a new State-Of-The-Art result on the RuCoLA leader board in 2022, reaching the Human Benchmark of that time for *accuracy* [4] [5].

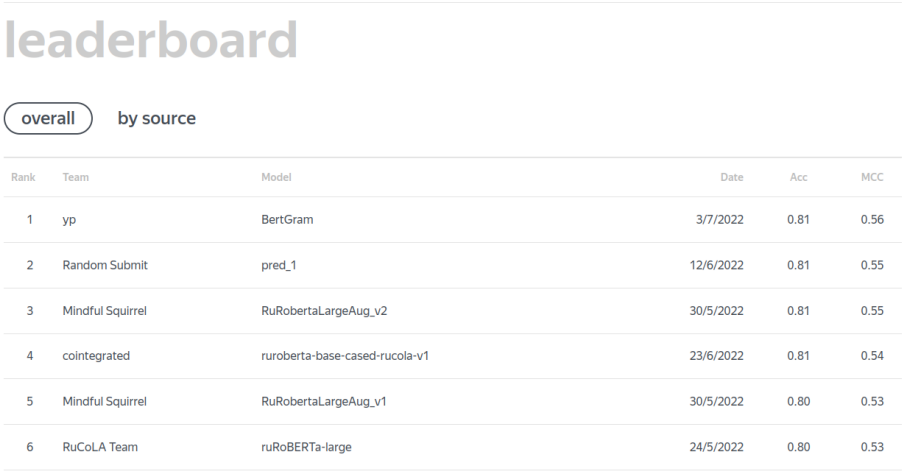

Figure 4: The model got a SOTA result on the RuCoLA leaderboard

Unfortunately, it didn't manage to repeat the success this time resulting in only 7th position on current leaderboard. The dataset collected and prepared for training the grammemes model was thrice as bigger this time. Ironically, it lead to GPU quota exhaustion on Google.Colab, so the model could not train properly. Same applies to hyper-parameter selection and main model training.

leaderboard

overall | by source

Rank	Team	Model	Date	Expert		Machine	
				Acc	MCC	Acc	MCC
	RuCoLA Team	Human Benchmark	24/5/2022	0.84	0.57	TBA	TBA
1	yp	BertGram	3/7/2022	0.84	0.54	0.80	0.55
2	Random Submit	pred_1	12/6/2022	0.83	0.52	0.80	0.56
3	Mindful Squirrel	RuRobertaLargeAug_v2	30/5/2022	0.83	0.52	0.80	0.55
4	cointegrated	ruroberta-base-cased-rucola-v1	23/6/2022	0.84	0.54	0.79	0.53
5	Mindful Squirrel	RuRobertaLargeAug_v1	30/5/2022	0.83	0.51	0.79	0.53
6	RuCoLA Team	ruRoBERTa-large	24/5/2022	0.83	0.51	0.79	0.53

Figure 5: The model got a SOTA result on the RuCoLA leaderboard

6.1 Ideas for improvement

- the original BERT from the Transformers library does a rather primitive Pooling of hidden states (takes the hidden state related to the first token in the sequence); applying another pooling may improve the model;
- perhaps we can apply a similar approach to improve quality on semantic errors – there are corpora with semantic markup of texts (for example, from the National Corpus of the Russian language `https://ruscorpora.ru/page/instruction-semantic`).

.

7 Conclusion

We managed to collect a dataset for training a model used later for transfer learning, using two different manually marked-up corpora. Unfortunately, we didn't managed to use the entire dataset for training due to lack of computational resources. However, the model was successfully trained on a part of that dataset. Using that model and pre-trained Transformer-based model we designed and trained a model for assessing the grammatical acceptability of sentences in Russian, which takes grammar features into account.

References

[Bahdanau et al., 2014] Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate.

[Batra et al., 2021] Batra, S., Jain, S., Heidari, P., Arun, A., Youngs, C., Li, X., Donmez, P., Mei, S., Kuo, S., Bhardwaj, V., Kumar, A., and White, M. (2021). Building adaptive acceptability classifiers for neural NLG. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 682–697, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

[Chung et al., 2020] Chung, H. W., Févry, T., Tsai, H., Johnson, M., and Ruder, S. (2020). Rethinking embedding coupling in pre-trained language models.

[Conneau et al., 2019] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale.

[Devlin et al., 2018] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding.

[Gusev, 2021] Gusev, I. (2021). rnnmorph.

[Korobov, 2015] Korobov, M. (2015). Morphological analyzer and generator for russian and ukrainian languages. In Khachay, M. Y., Konstantinova, N., Panchenko, A., Ignatov, D. I., and Labunets, V. G., editors, *Analysis of Images, Social Networks and Texts*, volume 542 of *Communications in Computer and Information Science*, pages 320–332. Springer International Publishing.

[Liu et al., 2019] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach.

[Raffel et al., 2019] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need.

[Wolf et al., 2020] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on*

Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online. Association for Computational Linguistics.