In Case of Russian - BERT the Noun

David Djambazov
University of California, Berkeley
djambazov@berkeley.edu

Abstract

This paper explores the possibility of applying multilingual BERT Transformers to Grammar Error Correction of Russian cases. BERT-based morphological taggers have had excellent success identifying cases in correct sentences in Russian. However, I find that they struggle to deal with erroneous examples as information about the appropriate case is encoded not just in the noun form, but more importantly in the context, including verbs and prepositions. That led me to a BERT-based error-classification approach on a synthetic dataset of erroneous examples. The results demonstrate that case-pertinent information is encoded in pre-trained BERT models and can be harnessed for the task of building a case-error correction system.

1 Introduction

Grammar Error Correction (GEC) has been seen traditionally as an important tool for language learners and non-native communicators. Yet, despite the recent progress of GEC for English, other languages have lagged. The rapid rise of Transformer-architecture models [Vaswani and et al., 2017], pre-trained on large multilingual corpora, such as multilingual BERT (M-BERT) [Devlin and et al., 2019], has opened new avenues for developing interesting language-specific approaches for solving pertinent non-English GEC problems.

This paper looks at the very specific problem of Russian grammatical cases with a view to gauging the feasibility of building an automatic case correcting system. There are two potential starting points for exploration. One utilizes fine-tuned morphological taggers, capable of providing morphological annotations including case classification with high accuracy. The other is to use *probing* techniques to extract the information about cases encoded in the pre-trained M-BERT model [Goldberg, 2019, Mikhailov and et al., 2021]. As an extension to the second approach I also look at attention maps across heads and layers, as described in literature [Rogers and et al., 2019, Vig and Belinkov, 2019], to see where we could seek cues to the identity of individual words with erroneous cases. The conclusion is that a combination of approaches gives the most intriguing direction.

2 Short Linguistic Backstory

Russian is a highly inflectional, morphologically rich language with nouns, pronouns, adjectives and numerals subject to change (declension) in different cases. The Nominal declension has six cases in two numbers (plural and singular) and three grammatical genders. This provides great structural flexibility as noun cases can supplant the use of prepositions altogether. Articles do not exist at all in Russian. Their meaning is conveyed by a combination of contextual and case-related cues. The trade-off is inflectional complexity. In addition to nouns generally changing form depending on case and number, there are three distinct declensions mostly depending on gender and word-endings. Within a sentence, subject nouns are always in the Nominative case, while for the rest, selection is governed by the choice of verb and the question to which the noun is the answer. For example an answer to ”Where is something occurring?” is always in the Prepositional case. Prepositions, to the extent they are used, work only with specific, though not necessarily unique, cases. Dependent adjective and pronoun case agreement with the relevant nouns is absolute. For the purposes of this projects I consider the six main cases of Russian. There are several others, but they are considered archaic and are extremely rare in any modern text.

3 Project Parameters

The main motivation of this work is to work toward a tool that can take in a Russian sentence, check for case errors and return the right forms of all nominals. Other forms of error detection and correction (such as spelling, verb conjugation and number agreement, prepositions, etc.) are beyond the scope of the project.

3.1 Goals

For this paper, the starting goal is to build a model that can distinguish between a correct sentence in Russian and one with case errors. Next is to take a sentence in Russian that has a single case error and with high precision classify the error and the correct case. Final step is to propose an approach for the identification of the errant noun phrase and substitution with the correct form.

3.2 Related Work

Most of the prior relevant work has been focused on the task of morphological tagging of correct Russian sentences. Comparison tables across languages show that this task for Russian can been accomplished to a very high-degree even with pre-Transformer approaches [Straka, 2018] such as character-based bi-directional LSTMs [Heigold and et al., 2017]. More recently, the team at DeepPavlov [Burtsev and et al., 2018] has improved on those results with a BERT-based model trained on Universal Dependencies corpora (version 2.3). Specifically, their model is built on top of the monolingual extension of M-BERT known as RuBERT [Kuratov and Arkhipov, 2019].

There's been relatively less focus on more pertinent GEC tasks for Russian. One approach [Rozovskaya and Roth, 2019] has been to train a phrase-based MT system [Susanto and et al., 2014] for GEC. The results for case correction are less than 40% on both precision and recall, which is not all surprising, given the complexity of their task (GEC on multiple types of errors), the limited data set and the MT modelling choice.

Another approach has been proposed in conjunction to a recent effort to create a suite of morphological probing tasks based on the detection of guided sentence perturbations [Mikhailov and et al., 2021]. The models employed are the main multilingual transformer models released as a part of HuggingFace library including M-BERT. In the probing task for cases (identifying cases in correct sentences), they improve on the DeepPavlov results from 97% to 99%. The case perturbation task focuses on altering the case just for the subject noun – a task easier task than the one proposed in the current paper as subject nouns should always be in the Nominative case. Never-the-less, as they report a result of 79%, it seems that the Transformer-based approach is quite promising.

Perturbed Noun Case Tag		
Adjective change	Same case	Nom. case
No	36.87%	39.61%
Yes	13.71%	45.25%

Table 1: Morphological tagger struggles with single-noun perturbations of 584 sentences from singular non-Nominative case to Nominative case.

3.3 Methods

I first looked to see how the high precision of the morphological tagging models for correct sentences can be best utilized for incorrect sentences. One idea is to use M-BERT as a masked language model and then determine the case of the top masked word predictions. As it turned out, the problem with this method approach is that it relies on detection of differences between tag of the original word and the tag of the predicted word. That, however is easier said than done. On one hand, simply using the word form to determine the case can be problematic due to disambiguation complexities as forms can be identical in different cases and declensions. On another, Russian grammatical structure provides several other clues as to the identity of the noun case, including verbs, prepositions, pronoun and adjective agreement. Unsurprisingly, morphological tagging models step on all of those to make their case determination [Mikhailov and et al., 2021]. The net result is that just changing the word form is often not enough of a perturbation to produce a different case tag. Table 1 shows the level of mismatch between perturbed cases and their tags even for a simple change to Nominative singular case. Importantly, for nouns with an adjective in the preceding position, changing the adjective as well can help reduce the tendency of the tagger to report the unperturbed case. Still, the success in correctly designating the noun form as Nominative is barely at 45%.

Another approach is to see if BERT, as a Masked Language Model (MLM), could give a higher probability to the specific correct form of the noun under question. Here we again run into difficulty as wordpiece tokenization means that different word forms are often broken up into different number of tokens. The stem tokens are usually the same, but there's no guarantee that the end tokens correspond exactly to the case inflections. There are even examples where the word forms of the same noun in different cases are broken up in completely different ways: *dev* + *##ushka* versus *dev* + *##ush* + *##ke*.

The next logical idea is to let the tagger do what it does best – tag correct sentences, while approaching the perturbation and error detection from a different direction. If we could take a large, high-quality corpus of Russian texts, tag the cases for each sentence and then construct a dictionary of word forms that covers as many of the variations according to case, number and gender (for adjectives) as are present in the corpus, we can easily assemble a large synthetic perturbed dataset from the original correct sentences. Thus we would know to a high degree of precision both the proper case and the case corresponding to the perturbed noun form. The final step is to then fine-tune BERT for a straightforward classification task with 31 classes (1 negative class for correct sentences and 30 positive ones, corresponding to a full complement of shifts among the 6 cases). After that point, identifying the positions of the wrong cases would be last piece of the puzzle before completing an end-to-end case correction system for single-noun errors. I look at attention maps for clues.

3.4 Data

I use the Meduza corpus - a collection of articles published on the *meduza.io* platform based out of Lithuania and featuring some Russia's best long-form journalists. The corpus contains 70,000 texts, comprised of 1.39 million sentences or 17.5 million tokens. Tagging and extracting cased nouns and adjectives from the dataset yielded a dictionary of 63,000 distinct base words and 318,000 word forms. The sentences were then perturbed with the goal of yielding 10,000 examples for each of the 30 possible case shifts. In the final count, three somewhat overlapping datasets were constructed. One for baseline modelling with 600,000 sentences with 46:54 split between correct and perturbed sentences. A binary M-BERT classification set with 280,000 sentences with a split 44:56. The third dataset included 275,000 perturbed sentences and 10,000 correct ones for multi-class classification. Due of the frequency difference between some cases such as Nominative (present in virtually every sentence) and others, such as Dative, we would need to perturb a much higher number of sentences to achieve a perfectly balanced distribution. That might not even be a desirable characteristic, as it would make training and validation class distributions differ significantly from what would naturally occur in texts. For that reason, I chose a configuration where the the least represented case shift (Dat to Loc) is about 50% of the example count target.

4 Experimental Setup

4.1 Models

At the focal point of the experimentation was multilingual base cased BERT pre-trained model. Despite some evidence that other models, such as XLM-R [Conneau and et al., 2019], or monolingual extensions of M-BERT such as RuBERT might perform better at certain tasks, probing out-of-the-box BERT's ability to perform a complex error detection task for a language other than English can offer valuable information for researchers looking to explore pre-trained BERT's usefulness for lower-resource language tasks.

4.2 Metrics

While both precision and recall are interesting for this setup, depending on the business case there are two divergent targets. For language learning purposes precision is key as many false positives can lead to learner confusion as they struggle to gain intuition about what are mistakes and what aren't. In the fast-paced environment of business communication, minimizing the final number of errors is most important. In this case, accuracy might be the better measure. When designing a correction system, introducing Type 2 errors (suggesting changing correct sentences) is probably a lot less desirable, so that's closer to the former goal and thus weighing precision heavier makes sense.

4.3 Baseline

In the binary classification task, the lowest possible baseline is majority class, which in this case is 54% correct sentences. I took the approach of training custom Russian *word2vec* 200-dimensional embeddings from the corpus using the *gensim* library and then feed them to a simple Convolutional Neural Network (CNN) and try different filters sizes to see if vector representations of the local context capture some case-relevant information to help with classification. The architecture is a straight-forward ReLU-activated convolutional layer with 100 filters of size 5 to hopefully capture enough context to predict the right case. Following is a dropout of 0.5, a max-pooling layer and a 100-neuron hidden layer before a sigmoid-activated classification layer with 1 neuron.

4.4 Binary Classification with M-BERT

The setup of this classifier is done via a single 200-neuron hidden layer taking the sentence [CLS] token from the underlying BERT model, applying dropout and feeding the output to a softmax-activated classification layer to get the designation. After some hyper-parameter tuning on a smaller dataset, I chose to go with a learning rate of 2×10^{-5} and a dropout rate of 0.2. Freezing layers was also tested, but ultimately I found that training M-BERT's layers is helpful as it reads through the high-quality Russian dataset. The ratio of training to validation is 80/20.

4.5 Multi-class Classification with M-BERT

For this model the key changes from the binary iteration is that the classification layer has 31 neurons corresponding to the number of classes and that it is trained on almost twice as many perturbed sentences.

5 Results

5.1 Baseline

The *word2vec* representation approach seemed like a reasonably interesting idea. As we see in Figure 1, word forms are clustered close to each other, but there's still some separation, as semantically relevant words are also quite close.

After 150 epochs the model achieved a training loss of and a testing loss of 0.5747 and testing loss of 0.6079. The accuracy for the test set was 0.6719 - an improvement over majority class and a reasonably looking baseline. I also ran some combinations of filter sizes and quantities, but the simple model was as good as it got.

5.2 Binary M-BERT classification

Turning to the heart of the question: has base multilingual BERT learned something about Russian cases in pre-training and encoded it in its 144 heads across 12 layers. The answer is emphatically yes.

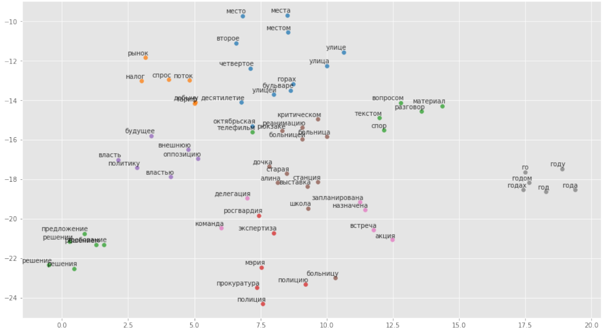


Figure 1: Word2vec embeddings: Case declension word forms are close, but so are more semantically relevant nouns and adjectives.

Binary Classification Table			
	Precision	Recall	F_1
Wrong Cases	0.93	0.94	0.94
Correct Cases	0.92	0.92	0.92
accuracy			0.93
macro avg	0.93	0.93	0.93
weighted avg	0.93	0.93	0.93

Table 2: The binary classifier based on fine-tuned M-BERT Transformer does well to pick out the sentences with the wrong cases.

Table 2 shows that all the important metrics are consistent and show that the model has done well to separate the correct sentences from the perturbed ones. Breakdown of mislabeled examples by original and shifted cases (Figure 2) paints an even clearer picture. Certain case perturbations lead to much higher error rates than others. One example would be Accusative to Genitive (such as **oboznacheniE** versus **oboznacheniYA**). This is an interesting avenue for further research.

5.3 Multiclass M-BERT model

The overall accuracy is 0.78. That's less impressive than what we got for the binary classification, but perhaps we can somehow improve it. Figure 3 shows the confusion matrix of the 31 classes. We see that they are neatly lined on or along the diagonal and mostly above 80%. Something else is intriguing. To the extent that there are some more prominent off-diagonal elements, they are right next door. This led me to the following idea. Classifying the exact case shift (e.g. Loc to Dat) is all fine and good, but for the purpose of error correction, the most important steps are to correctly determine that there has been an error in case (something we

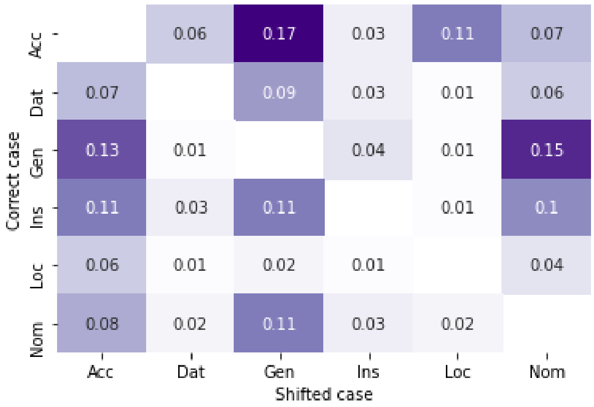


Figure 2: Binary M-BERT model: Some case shifts are much more prone to misclassification.

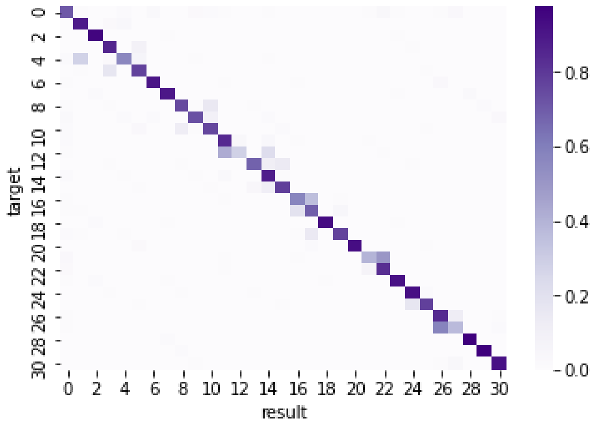


Figure 3: Multi-class model: BERT recognizes what different case mistakes look like.

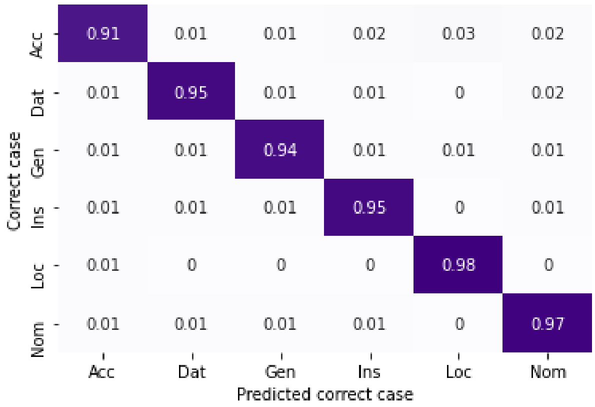
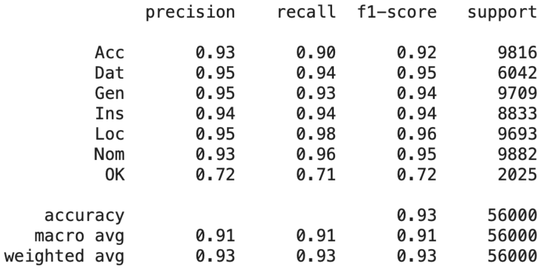


Figure 4: Multi-class M-BERT: Ya tebya slyshu (I can hear you) - the original case signal is strong.



```
              precision    recall  f1-score   support

         Acc       0.93      0.90      0.92      9816
         Dat       0.95      0.94      0.95      6042
         Gen       0.95      0.93      0.94      9709
         Ins       0.94      0.94      0.94      8833
         Loc       0.95      0.98      0.96      9693
         Nom       0.93      0.96      0.95      9882
          OK       0.72      0.71      0.72      2025

    accuracy                           0.93     56000
   macro avg       0.91      0.91      0.91     56000
weighted avg       0.93      0.93      0.93     56000
```

Figure 5: Multi-class M-BERT: Performance on par with the binary classification.

showed to be feasible in the binary classification) and what the correct case should be. Since each predicted case shift also uniquely predicts a case of origin, we could look at the results just from that perspective. That is, it is almost all the same if the model got the destination case wrong as long as the original is right. Let's take a look at a confusion matrix for the true positives (all the perturbed examples that were classified as such, and see if the true and predicted original cases match).

The result from Figure 4 is almost unexpectedly good. The initially low recall numbers for some classes were mostly due to the model missclassifying the destination case, but not the origin. It is almost like having a native speaker, who hasn't studied much grammar in school and can't identify cases by word forms alone, but can very much utter a correct sentence because it sounds right.

Figure 5 shows the recast classification report. We note that the precision and recall for the correct examples are just above 70%, but the support is also quite low, as the original cases now aggregate several case shifts. It should be looked into further, but here we have our evidence of the viability of a case correction system in the single-noun per sentence error scenario.

6 Attention analysis

In pursuit of the last element of a viable case corrector I looked at attention maps to assess the possibility of identifying the position of the case mistake so successfully picked up by the BERT-based classifier.

Colloquially known as Bertology, the analysis of attention patterns across Transformer heads and layers has been an exciting area of recent development [Rogers and et al., 2020]. I studied several other papers [Tenney and et al., 2019, Rogers and et al., 2019, Clark and et al., 2019] and found the code from Vig and Belinkov [2019] particularly helpful. One idea worth exploring would be to look for systemic patterns across pairs of sentences with correct and errant cases and try to understand if attention maps highlight consistently information about where the error might be. Here is an interesting example in Figure 6 and Figure 7.

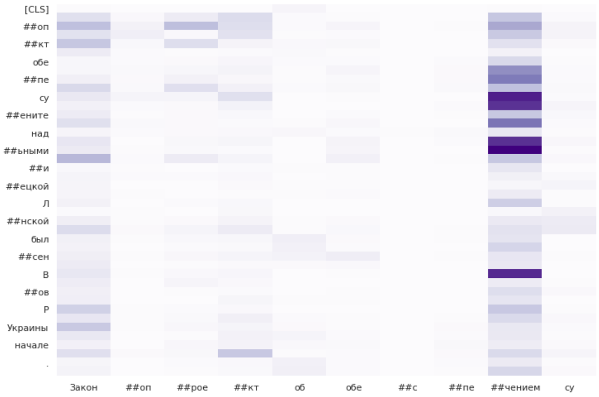


Figure 6: Attention: Wrong case (Ins) token is attended differently across the entire sentence, but in particular by inflection tokens of other nouns and adjectives.

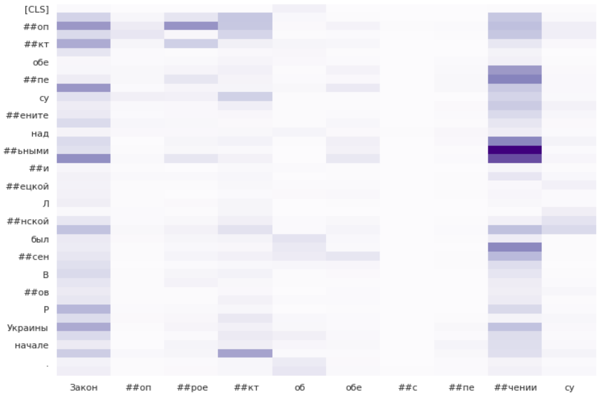


Figure 7: Attention: Correct case (Loc) token.

7 Conclusion

This paper shows good progress in the stated goal. Some useful tricks bring to bear the power of Transformers to one of the major hurdles for non-native Russian learners and communicators - case choice and declensions. We're not quite at the end of the journey, but the road is clear. Further work can focus on comparing the relative performance of different multi and monolingual Transformer models, on error analysis of specific case shifts, and not least on the performance of this approach for sentences with multiple case errors. Finally, a thorough analysis of the mechanisms employed by the model to achieve good results could open the way to even better solutions.

References

Burtsev and et al. Deeppavlov: Open-source library for dialogue systems”. In *Proceedings of ACL 2018, System Demonstrations*, pages 122–127, 2018.

K. Clark and et al. What does bert look at? an analysis of berts attention. *arXiv preprint*, 2019.

A. Conneau and et al. Unsupervised cross-lingual representation learning at scale. *arXiv preprint*, 2019.

J. Devlin and et al. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 4171–4186, 2019.

Y. Goldberg. Assessing berts syntactic abilities. *arXiv preprint*, 2019.

G. Heigold and et al. An extensive empirical evaluation of character-based morphological tagging for 14 languages. 2017.

Y. Kuratov and M. Arkhipov. Adaptation of deep bidirectional multilingual transformers for russian language. *arXiv preprint*, 2019.

V. Mikhailov and et al. Morph call: Probing morphosyntactic content of multilingual transformers. *arXiv preprint*, 2021.

A. Rogers and et al. Revealing the dark secrets of bert. 2019.

A. Rogers and et al. A primer in bertology: What we know about how bert works. *arXiv preprint*, 2020.

A. Rozovskaya and D. Roth. Grammar error correction in morphologically rich languages: The case of russian. *Transactions of the Association for Computational Linguistics*, 7:1–17, 2019.

M. Straka. Udpipe 2.0 prototype at conll 2018 ud shared task. 2018.

H. Susanto and et al. System combination for grammatical error correction. 2014.

I. Tenney and et al. Bert rediscovers the classical nlp pipeline. In *Proceedings of the Conference of the Association for Computational Linguistics, ACL*, pages 4593–4601, 2019. URL http://arxiv.org/abs/1905.05950.

A. Vaswani and et al. Attention is all you need. *Advances in neural information processing systems*, pages 5998–6008, 2017.

J. Vig and Y. Belinkov. Analyzing the structure of attention in a transformer language model. 2019.