Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2025"

April 23–25, 2025

BERT-like Models for Automatic Morpheme Segmentation of the Russian Language

Dmitry Morozov
Russian National Corpus &
Novosibirsk State University
Novosibirsk, Russia
morozowdm@gmail.com

Anna Glazkova
Russian National Corpus &
University of Tyumen
Tyumen, Russia
a.v.glazkova@utmn.ru

Timur Garipov
Russian National Corpus &
Novosibirsk State University
Novosibirsk, Russia
garipov154@yandex.ru

Abstract

Current approaches to automatic morpheme segmentation for the Russian language rely on machine learning, primarily neural network methods. Among the architectures presented, the best results have been achieved using convolutional neural networks and LSTM networks. However, the quality of automatic annotation is far from ideal, especially when dealing with roots that were not present in the training dataset. In this work, we present a new approach to morpheme segmentation based on fine-tuning BERT-like models. Through comparisons using two morpheme dictionaries with different segmentation paradigms, we demonstrated the superiority of our approach over previous ones, including when working with unfamiliar roots. The best result was achieved by fine-tuning the RuRoBERTa-large model: when working with random words, the share of completely correct segmentations increased from 88.5-90.8% to 92.5-93.5%, and when working with unfamiliar roots, it improved from 70.5-72.6% to 74.9-77.2%. Error analysis of the model showed that root nests not encountered in the training dataset can be distributed into two groups during testing: "recognizable", meaning those for which more than 90% of the words are correctly analyzed, and "unknown", meaning those for which the proportion of correct segmentations is less than 10%.

Keywords: automatic morpheme segmentation, Russian language morphology, machine learning, BERT
DOI: 10.28995/2075-7182-2025-23-XX-XX

BERT-подобные модели для автоматического построения морфемных разборов слов русского языка

Дмитрий Морозов
НКРЯ & НГУ
Новосибирск, Россия
morozowdm@gmail.com

Анна Глазкова
НКРЯ & ТюмГУ
Тюмень, Россия
a.v.glazkova@utmn.ru

Тимур Гарипов
НКРЯ & НГУ
Новосибирск, Россия
garipov154@yandex.ru

Аннотация

Актуальные подходы к автоматическому построению морфемных разборов для русского языка опираются на машинное обучение и, в первую очередь, нейросетевые подходы. Среди представленных архитектур лучших результатов удалось добиться при помощи свёрточных нейронных сетей и LSTM-сетей. Тем не менее, качество автоматической разметки далеко от идеального, в особенности, при работе с корнями, не встретившимися в обучающей выборке. В настоящей работе мы представляем новый подход к морфемной сегментации на базе дообучения BERT-подобных моделей. В ходе сравнения на материале двух морфемных словарей с различной парадигмой членения мы продемонстрировали превосходство нашего подхода над предыдущими, в том числе, при работе с незнакомыми корнями. Лучший результат был достигнут при помощи дообучения модели RuRoBERTa-large: при работе со случайными словами качество разметки выросло с 88,5-90,8% до 92,5-93,5%, при работе с незнакомыми корнями — с 70,5-72,6% до 74,9-77,2%. Анализ ошибок модели показал, что корневые гнёзда, не встретившиеся в обучающей выборке, при тестировании распределяются на две группы: «узнаваемые», то есть такие, для которых более 90% слов разобрано корректно, и «неизвестные», то есть такие, для которых доля корректных разборов составляет менее 10%.

Ключевые слова: автоматическое построение морфемных разборов, морфемика русского языка, машинное обучение, BERT

1 Introduction

Morpheme segmentation refers to the division of a word into several morpheme substrings. Morpheme segmentation is one of the key aspects of studying the morphology of the Russian language, allowing the analysis of how words are formed from minimal meaningful units — roots, prefixes, suffixes, and others. This enables a deeper understanding of the structure of words, revealing patterns in their formation and changes. It is also in demand in the teaching of the Russian language, as many spelling rules rely on the ability of the speaker to identify individual morphemes within a word (Kisselev et al., 2024). Morpheme segmentation also shows promise as a subword tokenization strategy for language models. Recent studies (e.g., (Matthews et al., 2018; Nzeyimana and Niyongabo Rubungo, 2022)) demonstrate that morpheme-based tokenizers can outperform standard Byte-Pair Encoding (BPE) (Gage, 1994) in certain scenarios, yielding measurable improvements in model performance. Finally, morpheme annotation is used in large text corpora (Savchuk et al., 2024), which are employed for linguistic research, to enhance user search capabilities.

Since morpheme segmentation can sometimes be a complex task requiring the attention of professional linguists, morpheme dictionaries are created for practical purposes, such as the Word Formation Dictionary of the Russian Language (Tikhonov, 1990) and the Dictionary of Morphemes of the Russian Language (Kuznetsova and Efremova, 1986). However, it cannot be claimed that such dictionaries are comprehensive. Firstly, morpheme dictionaries typically include segmentations only for lemmata, which is not an ideal solution for the Russian language given its rich morphological possibilities. Secondly, the largest existing morpheme dictionaries contain around 100,000 to 150,000 lemmata, which is significantly smaller than the overall volume of the Russian language; for example, the Main Corpus of the Russian National Corpus (Savchuk et al., 2024) contains 200,000 to 250,000 different lemmata. A further complication arises from the differences in the paradigms of morpheme segmentation used in these dictionaries. Some dictionaries, such as the Word Formation Dictionary of the Russian Language, focus on the transparency of relationships and the presence of word formation chains in modern Russian, while others, like the Dictionary of Morphemes of the Russian Language, pay more attention to the etymology of words and their correspondence with other lexemes of similar structure. As a result, segmentations of the same lemma often differ across dictionaries. Additionally, authors may locally make decisions that contradict the paradigms they declare, complicating the automatic updating of dictionaries, while manual updates require significant regular investments of time and resources due to the emergence of new words in the language.

In this context, approaches based on machine learning may serve as suitable tools. A number of studies have previously been published on the Russian language (Sorokin and Kravtsova, 2018; Bolshakova and Sapin, 2019b; Bolshakova and Sapin, 2019a; Bolshakova and Sapin, 2022; Garipov et al., 2024; Morozov et al., 2024), which examine gradient boosting over decision trees, convolutional neural networks (CNN), and LSTM networks, demonstrating high quality in the generated segmentations. At the same time, it was shown (Morozov et al., 2024) that many presented algorithms perform significantly worse with words containing roots that did not appear in the training dataset: about 90% of fully correct segmentations for random words and around 70% for words with unfamiliar roots. Moreover, it has been demonstrated (Sorokin, 2022) that use of embeddings from pre-trained BERT-like models can improve the performance of algorithms, including on data with unfamiliar roots. Although BERT was used in these works only as a source of embeddings for the convolutional network, Peters and Martins showed (Peters and Martins, 2022) that subword transformers can show good results in the task of morpheme segmentation, including on the material of the Russian language. Thus, this raises a question: will the fine-tuning of pre-trained Transformer-based models be effective in the task of morpheme segmentation? In this paper, we build upon this idea and present a family of fine-tuned BERT-like models for the automatic morpheme segmentation of the Russian language lemmata, which surpass all previously presented approaches. Our models are available for download at `https://ruscorpora.ru/license-content/neuromodels`. The increase in the proportion of fully correct segmentations reached up to 3% for random words and up to 4.5% for words with unfamiliar roots.

2 Related Work

Morpheme segmentation algorithms can be classified according to two criteria. The first of these criteria is the type of segmentation. Two variants of segmentation are distinguished: surface and canonical (Cotterell et al., 2016). The distinction of canonical segmentation lies in the fact that it not only involves breaking a word into morphemes but also reconstructing the original form of the morpheme. For example, for the word "funniest" the surface segmentation is *funn-i-est*, while the canonical segmentation is *fun-y-est*. The second criterion is whether the algorithm's tasks include determining the types of morphemes. Most studies on the Russian language focus on surface segmentation with the identification of morpheme types. A notable exception is the SIGMORPHON 2022 competition (Batsuren et al., 2022), in which the task of canonical segmentation was addressed without determining the type.

The most studied architecture for the Russian language is an ensemble of convolutional neural networks, first presented in the work (Sorokin and Kravtsova, 2018). This model operates in a character-level BMES annotation format, where each letter is assigned a label indicating its position within the morpheme (B for beginning, M for middle, E for end, and S for single letter morpheme) and a label corresponding to one of seven morpheme types: ROOT, PREF (prefix), SUFF (suffix), END (ending), POST (postfix), LINK (linking vowel), HYPH (hyphen). The quality of the model is measured using F-score, precision, and recall for morpheme boundaries, character-level accuracy, and the proportion of fully correct segmentations. Based on the dataset relying on the Word Formation Dictionary of the Russian Language (Tikhonov, 1990), the authors achieve approximately 87% fully correct segmentations and about 96% correctly labeled letters. Within a series of papers (Bolshakova and Sapin, 2019a; Bolshakova and Sapin, 2019b), this architecture was compared with models based on gradient boosting over decision trees and with LSTM networks using two datasets: the dataset from the original paper and the CrossLexica dictionary (Bolshakov, 2013). The comparison showed a slight advantage of the ensemble of three LSTM networks over a similar ensemble of convolutional neural networks. An alternative approach is the use of generative models. For example, at the SIGMORPHON 2022 competition, the subword transformer DeepSPIN-3 (Peters and Martins, 2022) demonstrated the best performance across multiple languages, including Russian. In (Morozov et al., 2024), some approaches were compared using three datasets: one based on the Dictionary of Morphemes of the Russian Language (Kuznetsova and Efremova, 1986) and the two previously mentioned, with corrections made to the annotation errors of the morpheme types in the dataset based on the Word Formation Dictionary of the Russian Language. The best results were demonstrated by the ensemble of convolutional neural networks. Additionally, comparing the algorithm's results with expert evaluation indicated that when tested on random words, the algorithms are close to achieving expert-level performance.

At the same time, Morozov et al. (Morozov et al., 2024) studied a key drawback of existing approaches: due to the lack of semantic consideration when dealing with roots that did not appear in the training dataset, the quality of annotation sharply declines: around 90% fully correct segmentations for random words, about 70% for words with unfamiliar roots for CNN and LSTM approaches. For DeepSPIN-3 the situation is even worse: 80% correct segmentations on average and only 14% when working with unfamiliar roots. This shortcoming can be partially addressed by enriching the feature description with embeddings from a pre-trained BERT-like model (Sorokin, 2022; Morozov et al., 2024), but the improvement is relatively modest. Pranjic et al. (Pranjić et al., 2024) proposed a binary classifier based on the Glot500 model (Imani et al., 2023) that determines the presence of a morpheme boundary in a word at the character level. This method is difficult to consider practically significant due to the extremely high computational costs (for a word of length N, the model needs to be run N-1 times). At the same time, the potential of using pre-trained Transformer-based language models as an end-to-end solution remains insufficiently explored.

3 Datasets

We used two datasets, which were previously utilized in (Garipov et al., 2024; Morozov et al., 2024): Morphodict-T, based on the Word Formation Dictionary of the Russian Language (Tikhonov, 1990), and Morphodict-K, based on the Dictionary of Morphemes of the Russian Language (Kuznetsova and Efre-

	Morphodict-T	Morphodict-K
Unique words	95,895	75,649
Unique morphemes	15,899	8,079
Unique roots	15,253	7,148
Average morphemes per word	3.86	4.12
Average morpheme occurrence	23.29	38.56
Average root occurrence	7.54	12.24
Average root length in symbols	5.52	4.62

Table 1: Brief characteristics of the datasets

mova, 1986). Besides differences in the vocabulary, the source dictionaries rely on different paradigms of morpheme segmentation, resulting in a significantly higher fragmentation of morphemes in Morphodict-K. More detailed differences between the datasets are described, for example, in (Morozov et al., 2024). The use of the Morphodict-T dataset allows for better comparison of the obtained results with those from the previous works that utilized a similar dataset (Sorokin and Kravtsova, 2018; Bolshakova and Sapin, 2019a; Bolshakova and Sapin, 2019b). Morphodict-K, on the other hand, enables us to verify whether the results obtained are significantly related to the specifics of the particular dataset. This is particularly important in our case, as we do not claim to develop a universal morpheme segmentation algorithm, which is most likely impossible due to disagreements in defining morpheme segmentation itself (Iomdin, 2019). However, within this study, we investigate whether it is possible to automatically extend annotations from a specific morpheme dictionary to out-of-vocabulary words. Brief characteristics of Morphodict-T and Morphodict-K are provided in Table 1.

4 Models and Experimental Setup

We utilized two BERT-like models pre-trained for the Russian language: RuBERT-base-cased (Kuratov and Arkhipov, 2019) (hereinafter referred to as RuBERT) and RuRoBERTa-large (Zmitrovich et al., 2024) (hereinafter referred to as RuRoBERTa). The choice of these models is due to their well-established performance and application in a wide range of practical tasks. We framed the original task as a specific case of Named Entity Recognition: the model was provided with a sequence of letters forming a word, and the task of the model was to determine the class of each letter in the sequence according to the BMES annotation with types (an example is shown in Figure 1). We tested two options for forming input sequences. In the first case, we provided only the letters and their classes, effectively preventing the model's tokenizer from functioning. In the second case, the starting element of the sequence was the original lemma, which was assigned a special class of "0" (Figure 2). Our hypothesis was that having the complete lemma would allow the model to take into account the semantics of the word and improve its performance with roots that were not present in the morpheme training dataset but had been encountered by the model during the pre-training phase. These models are further designated by the index "lex".

We fine-tuned the models using two different methods for splitting the dataset: fully random split of the dataset and split-by-roots, where words with the same root could not appear in different folds (words with multiple roots were excluded from the dataset in advance). The first method of forming folds allows for an average assessment of the model's annotation quality, while the second one evaluates the annotation quality when dealing with unfamiliar roots. We applied 5-fold cross-validation for a more accurate assessment of the automatic labeling quality.

The models were fine-tuned for 30 epochs, using a learning rate of 4e-5 for RuBERT-base-cased model and 5e-6 for RuRoBERTa-large one. To assess quality, we used the metrics proposed in (Sorokin and Kravtsova, 2018) and supplemented in (Garipov et al., 2024): F-score, precision, and recall for morpheme boundaries (these metrics are further designated by the index "full"); F-score, precision, and

Figure 1: Example of input sequence

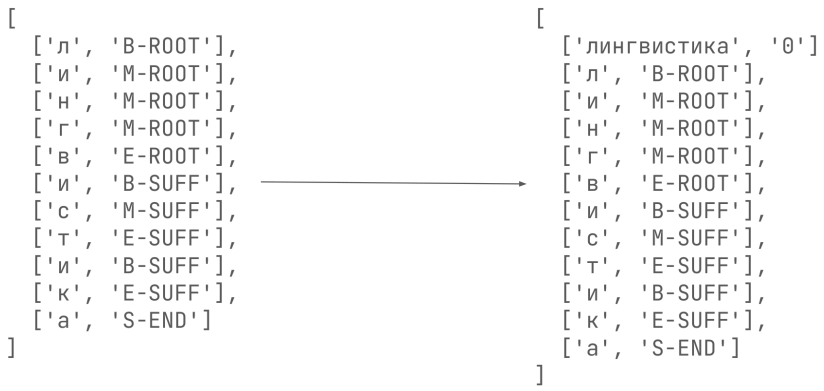

Figure 2: Example of adding lemma to input sequence

recall for root boundaries only (these metrics are further designated by the index “root”); as well as character-level accuracy (Accuracy) and word-level accuracy (WordAccuracy).

5 Results and Discussion

During the training process, we evaluated the quality of the models using word-level accuracy (WordAccuracy). In the case of the random split, the model quality changed significantly during the first few epochs: the difference between the first and tenth epochs is 15-20 percentage points. In contrast, for the split-by-roots mode, such an increase is not observed: the difference between the models after one epoch of training and after 30 epochs is only 2-3 percentage points. Several examples of the dependence of WordAccuracy on the number of epochs are shown in Figure 3.

Tables 2, 3 present the results obtained for the random split experiment using the Morphodict-T and Morphodict-K datasets, respectively, Tables 4 and 5 show results for the split-by-root experiment. In addition to the results for the fine-tuned models, the tables also include the results for the CNN ensemble as the best model from (Morozov et al., 2024). Each cell contains the average metric value measured during cross-validation, with the standard deviation indicated. The best metric value in each row is highlighted in gray.

The results obtained from the two datasets are largely similar. In the case of the random split, BERT-like models demonstrate higher quality than CNN ensemble, with RuRoBERTa-based models performing better than RuBERT-based ones. The superiority of RuRoBERTa models over the ensemble of convolutional neural networks is about 3 percentage points. For the RuBERT model, the one trained without using lemmata surpasses the $\text{RuBERT}_{\text{lex}}$ model in both Accuracy and WordAccuracy for both datasets.

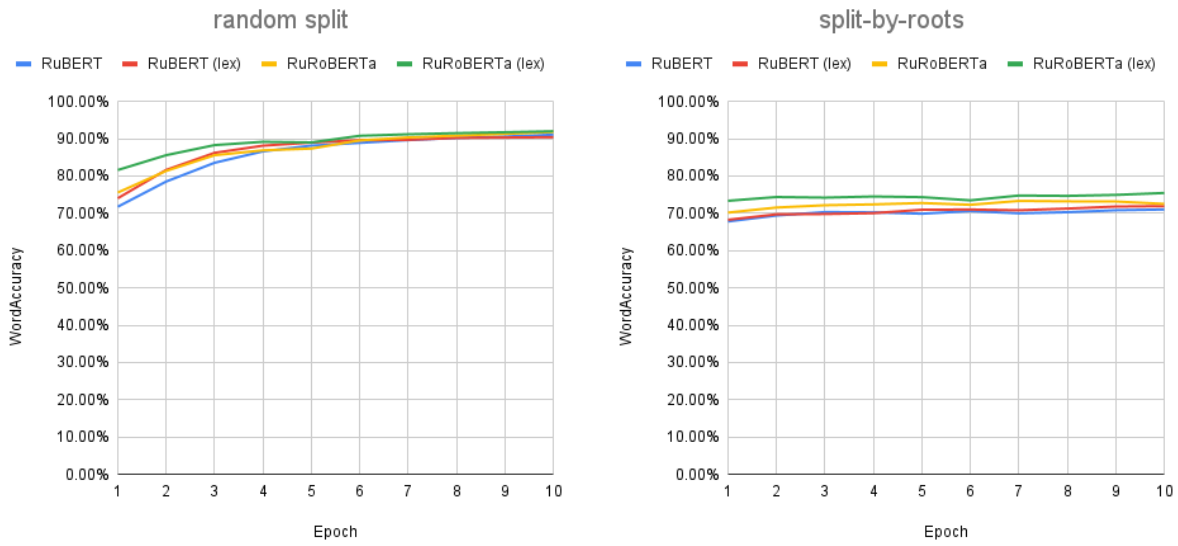

Figure 3: WordAccuracy of the models during the first ten epochs of fine-tuning on the Morphodict-K dataset

	RuBERT	$\text{RuBERT}_{\text{lex}}$	RuRoBERTa	$\text{RuRoBERTa}_{\text{lex}}$	CNN
$\text{Precision}_{\text{full}}$	98.60 ± 0.05	98.18 ± 0.11	98.57 ± 0.17	98.69 ± 0.03	97.79 ± 0.13
$\text{Recall}_{\text{full}}$	98.67 ± 0.05	98.32 ± 0.07	98.56 ± 0.18	98.84 ± 0.04	98.38 ± 0.06
$\text{F-score}_{\text{full}}$	98.63 ± 0.03	98.25 ± 0.07	98.56 ± 0.17	98.76 ± 0.02	98.09 ± 0.05
$\text{Precision}_{\text{root}}$	95.80 ± 0.11	94.61 ± 0.25	95.45 ± 0.44	96.13 ± 0.14	94.19 ± 0.29
$\text{Recall}_{\text{root}}$	95.57 ± 0.12	94.59 ± 0.20	95.18 ± 0.57	96.00 ± 0.08	93.99 ± 0.21
$\text{F-score}_{\text{root}}$	95.68 ± 0.11	94.60 ± 0.21	95.31 ± 0.50	96.07 ± 0.11	94.08 ± 0.25
Accuracy	97.59 ± 0.06	96.84 ± 0.11	97.39 ± 0.31	97.78 ± 0.05	96.61 ± 0.10
WordAccuracy	91.77 ± 0.19	89.59 ± 0.29	91.09 ± 1.01	92.47 ± 0.14	88.49 ± 0.28

Table 2: Results for the Morphodict-T dataset, random split

In the case of RuRoBERTa, this dependency does not hold: for Morphodict-T, $\text{RuRoBERTa}_{\text{lex}}$ performs better, while for Morphodict-K, the model qualities are nearly identical. When using the split by roots, the superiority of RuRoBERTa-based models over the others is confirmed, with $\text{RuRoBERTa}_{\text{lex}}$ outperforming the CNN ensemble by 4.5 percentage points across both datasets. It is also noteworthy that the drop in quality between the random split and the split by roots for models using lemmata is reduced: for CNN, the gap is about 18 percentage points, for $\text{RuBERT}_{\text{lex}}$ about 17.5 percentage points, and for $\text{RuRoBERTa}_{\text{lex}}$ between 16.5 and 17.5 percentage points. Performance in root extraction also improved, as evidenced by the F-score for root morphemes: for $\text{RuRoBERTa}_{\text{lex}}$, this metric is 1.5-2 percentage points higher than that of the CNN ensemble.

The $\text{RuRoBERTa}_{\text{lex}}$ models significantly outperform the CNN ensemble and RuBERT-based models. In comparison with RuRoBERTa models, $\text{RuRoBERTa}_{\text{lex}}$ models excel in the split-by-roots case, while in the random split case, they outperform for the Morphodict-T dataset and show nearly identical results for the Morphodict-K dataset. Therefore, the $\text{RuRoBERTa}_{\text{lex}}$ models should be considered the most promising in terms of practical application. Consequently, we conducted an analysis of the segmentations made by these models and compared them with the previous baseline, the CNN ensemble.

	RuBERT	$\text{RuBERT}_{\text{lex}}$	RuRoBERTa	$\text{RuRoBERTa}_{\text{lex}}$	CNN
$\text{Precision}_{\text{full}}$	99.00 ± 0.03	98.68 ± 0.09	99.05 ± 0.05	99.04 ± 0.06	98.58 ± 0.11
$\text{Recall}_{\text{full}}$	99.11 ± 0.03	98.81 ± 0.04	99.14 ± 0.03	99.17 ± 0.03	98.74 ± 0.09
$\text{F-score}_{\text{full}}$	99.05 ± 0.02	98.74 ± 0.03	99.09 ± 0.04	99.10 ± 0.04	98.66 ± 0.03
$\text{Precision}_{\text{root}}$	97.30 ± 0.11	96.15 ± 0.17	97.44 ± 0.20	97.37 ± 0.12	96.26 ± 0.13
$\text{Recall}_{\text{root}}$	97.24 ± 0.08	96.19 ± 0.18	97.43 ± 0.12	97.35 ± 0.08	96.22 ± 0.12
$\text{F-score}_{\text{root}}$	97.27 ± 0.08	96.17 ± 0.17	97.43 ± 0.16	97.36 ± 0.10	96.24 ± 0.11
Accuracy	98.10 ± 0.05	97.52 ± 0.08	98.19 ± 0.06	98.19 ± 0.06	97.40 ± 0.03
WordAccuracy	93.39 ± 0.14	91.51 ± 0.24	93.61 ± 0.16	93.54 ± 0.16	90.82 ± 0.13

Table 3: Results for the Morphodict-K dataset, random split

	RuBERT	$\text{RuBERT}_{\text{lex}}$	RuRoBERTa	$\text{RuRoBERTa}_{\text{lex}}$	CNN
$\text{Precision}_{\text{full}}$	95.27 ± 0.26	95.62 ± 0.42	94.67 ± 0.78	96.09 ± 0.34	94.46 ± 0.50
$\text{Recall}_{\text{full}}$	93.58 ± 0.50	94.33 ± 0.35	95.31 ± 0.90	95.08 ± 0.28	94.96 ± 0.45
$\text{F-score}_{\text{full}}$	94.42 ± 0.32	94.97 ± 0.32	94.98 ± 0.18	95.59 ± 0.22	94.71 ± 0.34
$\text{Precision}_{\text{root}}$	81.52 ± 0.58	82.47 ± 0.80	81.91 ± 0.48	84.37 ± 0.62	81.96 ± 0.73
$\text{Recall}_{\text{root}}$	81.47 ± 0.54	82.38 ± 0.84	82.26 ± 0.28	84.30 ± 0.61	81.98 ± 0.72
$\text{F-score}_{\text{root}}$	81.50 ± 0.56	82.42 ± 0.82	82.08 ± 0.34	84.34 ± 0.61	81.97 ± 0.72
Accuracy	89.58 ± 0.57	90.61 ± 0.52	90.15 ± 0.38	91.70 ± 0.34	90.16 ± 0.53
WordAccuracy	69.26 ± 1.76	72.12 ± 1.64	71.10 ± 1.36	74.95 ± 1.43	70.53 ± 1.79

Table 4: Results for the Morphodict-T dataset, split by roots

Despite the lower proportion of incorrect segmentations, the remaining annotation errors in the case of the random split exhibit a nature similar to the errors of the CNN ensemble (Morozov et al., 2024): insufficient consideration of semantics by the model and difficulties in distinguishing prefixes and roots in the case of borrowed morphemes. An example of the first type of error is the segmentation for the adverb "передом'' (*in front*): *пере:PREF/дом:ROOT*, whereas the correct segmentation is *пе-ред:ROOT/ом:SUFF*. It is important to note that in the Russian language, there is both the prefix 'пере-' (*over*) and the root '-дом-' (*house*), which is likely the cause of the error. An example of the second type is the segmentation of the word "примадонна" (*primadonna, diva*): although the morpheme boundaries are correctly annotated as *прима:ROOT/донн:ROOT/а:END* (in contrast to the CNN segmentation *при:PREF/мадонн:ROOT/а:END*), the type of the first morpheme in the Morphodict-K dataset is indicated as ***PREF***. When working with unfamiliar roots, compared to the CNN ensemble, the most notable change is the reduction in the number of cases with excessive separation of the prefix from the root, for example:

1. "кооперация" (*cooperation*)
   - Morphodict-K: *ко:PREF/опер:ROOT/а:SUFF/ц:SUFF/и:SUFF/я:END*
   - CNN: *ко:PREF/о:PREF/пер:ROOT/а:SUFF/ц:SUFF/и:SUFF/я:END*
   - $\text{RuRoBERTa}_{\text{lex}}$: *ко:PREF/опер:ROOT/а:SUFF/ц:SUFF/и:SUFF/я:END*
2. "огниво" (***flint***)
   - Morphodict-K: *огн:ROOT/ив:SUFF/о:END*
   - CNN: *о:PREF/гни:ROOT/в:SUFF/о:SUFF*
   - $\text{RuRoBERTa}_{\text{lex}}$: *огн:ROOT/ив:SUFF/о:END*

For more general conclusions, we randomly selected 50 errors made by $\text{RuRoBERTa}_{\text{lex}}$ for each dataset and each dataset split strategy. The analysis revealed that the most common errors were related to root extraction: either an unnecessary prefix or suffix is included in the root, or conversely, an excessive

	RuBERT	RuBERT$_{lex}$	RuRoBERTa	RuRoBERTa$_{lex}$	CNN
Precision$_{full}$	95.50 ± 0.27	95.69 ± 0.33	95.78 ± 0.43	96.38 ± 0.44	95.35 ± 0.33
Recall$_{full}$	94.72 ± 0.38	95.13 ± 0.35	94.74 ± 0.35	95.81 ± 0.26	95.04 ± 0.29
F-score$_{full}$	95.11 ± 0.25	95.41 ± 0.26	95.26 ± 0.36	96.09 ± 0.27	95.19 ± 0.27
Precision$_{root}$	81.84 ± 2.21	82.52 ± 2.16	82.37 ± 2.38	84.95 ± 2.01	82.46 ± 2.26
Recall$_{root}$	81.75 ± 2.13	82.67 ± 1.97	82.28 ± 2.47	84.78 ± 1.93	82.54 ± 2.25
F-score$_{root}$	81.79 ± 2.17	82.60 ± 2.06	82.32 ± 2.42	84.87 ± 1.97	82.50 ± 2.25
Accuracy	91.13 ± 0.42	91.67 ± 0.50	91.39 ± 0.62	92.82 ± 0.42	91.30 ± 0.47
WordAccuracy	72.46 ± 1.95	73.99 ± 1.71	73.32 ± 2.24	77.17 ± 1.96	72.63 ± 2.22

Table 5: Results for the Morphodict-K dataset, split by roots

morpheme is extracted from the root. In Morphodict-T, for the random split, there were 39 out of 50 errors regarding root boundaries, while for the split-by-roots, there were 47 out of 50. In Morphodict-K, for the random split, there were 27 out of 50 errors regarding root boundaries, and for the split-by-roots, there were 40 out of 50. The higher number of such errors in Morphodict-T can be attributed to the morpheme segmentation paradigm adopted in (Tikhonov, 1990): roots are longer and more diverse, and historical prefixes and suffixes are often included within the root, which the model fails to recognize. Among other errors, the overwhelming majority are related to the segmentation of suffixes (e.g., *-ныч-* vs *-н-ыч-*, *-ирова-* vs *-ир-ова-*, etc.). This is largely associated with inconsistent segmentation in the datasets themselves.

We aimed to verify whether the model is indeed capable of "recognizing" unfamiliar roots or if the distribution of errors among the roots is more random in nature. To do this, we calculated the proportion of correctly annotated words for each root from the test set for the models trained in the split-by-roots mode. The results are presented in the form of a histogram in Figure 4. It is easy to see that the distributions for both datasets are similar and allow us to conclude that for the majority of roots, the model either "recognized" the root and identifies it correctly almost always, or it "did not recognize" it and almost always makes errors. Intermediate cases are quite rare. The question of which specific roots the model "recognizes" and which it does not, and how they relate to the training dataset, merits further investigation, which we plan to conduct in the future.

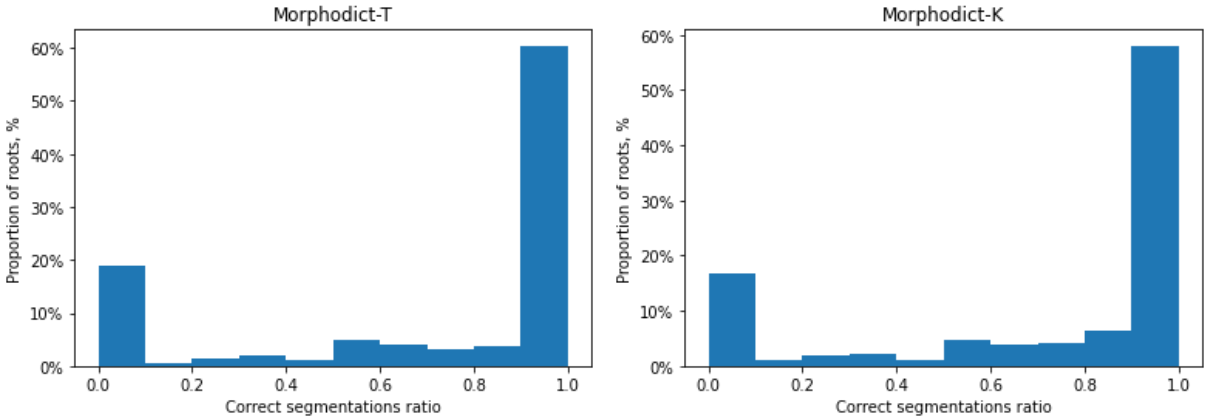

Figure 4: Distribution of the correct segmentations ratio for words containing a specific root

6 Conclusion

Automatic morpheme segmentation of Russian words is a relevant applied task in natural language processing. For a long time, the leading approaches to this task were ensembles of convolutional neural networks and LSTM networks. However, these approaches perform significantly worse with roots that were not present in the training dataset. In this work, we presented a family of fine-tuned BERT-like models for automatic morpheme segmentation that demonstrate better results compared to previous approaches. The model based on RuRoBERTa, trained for character-level annotation of words with the inclusion of the lemma in the input sequence, outperforms the convolutional neural network ensemble by 3% on random words and by 4.5% on words with unfamiliar roots. These results were obtained using two datasets with different paradigms of morpheme segmentation. Error analysis conducted on the best model shows that unfamiliar roots during testing are distributed into two categories: those that the model "recognized", meaning that the words are correctly analyzed more than 90% of the time, and those that the model "did not recognize", meaning that the words are correctly analyzed less than 10% of the time.

Two main limitations of our study should be highlighted. The first pertains to the complexity of training the model: one epoch of fine-tuning of RuRoBERTa on an NVIDIA RTX 4090 GPU took us about 2 minutes, whereas the ensemble of convolutional networks did not require a GPU at all and trained significantly faster. The second limitation relates to the potential application area of the models: the datasets used contain only the lemmata of words, while practical applications often require segmentations of word forms. Additionally, using word forms could enhance the quality of the final model. In the future, we plan to conduct a series of experiments to expand the existing datasets with segmentations of word forms and compare the quality of models trained solely on lemmata with those trained on the expanded data.

References

Khuyagbaatar Batsuren, Gábor Bella, Aryaman Arora, Viktor Martinovic, Kyle Gorman, Zdeněk Žabokrtský, Amarsanaa Ganbold, Šárka Dohnalová, Magda Ševčíková, Kateřina Pelegrinová, Fausto Giunchiglia, Ryan Cotterell, and Ekaterina Vylomova. 2022. The SIGMORPHON 2022 shared task on morpheme segmentation. // Garrett Nicolai and Eleanor Chodroff, *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, P 103–116, Seattle, Washington, July. Association for Computational Linguistics.

I.A. Bolshakov. 2013. Krossleksika: universum sviazi mezhdu russkimi slovami [crosslexica: a universe of links between russian words]. *Biznes-informatika*, №3 (25):12–19.

E. I. Bolshakova and A. S. Sapin. 2019a. Comparing models of morpheme analysis for Russian words based on machine learning. // *Komp'juternaja Lingvistika I Intellektual'nye Tehnologii*, volume 18, P 104 –113.

Elena Bolshakova and Alexander Sapin. 2019b. Bi-LSTM model for morpheme segmentation of Russian words. // Dmitry Ustalov, Andrey Filchenkov, and Lidia Pivovarova, *Artificial Intelligence and Natural Language*, P 151–160, Cham. Springer International Publishing.

Elena I. Bolshakova and Alexander S. Sapin. 2022. Building a combined morphological model for russian word forms. // Evgeny Burnaev, Dmitry I. Ignatov, Sergei Ivanov, Michael Khachay, Olessia Koltsova, Andrei Kutuzov, Sergei O. Kuznetsov, Natalia Loukachevitch, Amedeo Napoli, Alexander Panchenko, Panos M. Pardalos, Jari Saramäki, Andrey V. Savchenko, Evgenii Tsymbalov, and Elena Tutubalina, *Analysis of Images, Social Networks and Texts*, P 45–55, Cham. Springer International Publishing.

Ryan Cotterell, Tim Vieira, and Hinrich Schütze. 2016. A joint model of orthography and morphological segmentation. // Kevin Knight, Ani Nenkova, and Owen Rambow, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, P 664–669, San Diego, California, June. Association for Computational Linguistics.

Philip Gage. 1994. A new algorithm for data compression. *C Users J.*, 12(2):23–38, February.

Timur Garipov, Dmitry Morozov, and Anna Glazkova. 2024. Generalization ability of CNN-based Morpheme Segmentation. // *2023 Ivannikov Ispras Open Conference (ISPRAS)*, P 58–62.

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. // Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, P 1082–1117, Toronto, Canada, July. Association for Computational Linguistics.

B. L. Iomdin. 2019. How to define words with the same root? *Russian Speech = Russkaya Rech'*, (1):109–115, February.

Olesya Kisselev, Irina Dubinina, and Galina Paquette. 2024. A corpus-based study on orthographic errors of Russian heritage learners and their implications for linguistic research and language teaching. *Languages*, 9(4).

Y Kuratov and M Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for Russian language. // *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, P 333–339, Moscow, Russia, June.

A. I. Kuznetsova and T. F. Efremova. 1986. *Dictionary of Morphemes of the Russian Language*. Russkii yazyk, Moscow.

Austin Matthews, Graham Neubig, and Chris Dyer. 2018. Using morphological knowledge in open-vocabulary neural language models. // Marilyn Walker, Heng Ji, and Amanda Stent, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, P 1435–1445, New Orleans, Louisiana, June. Association for Computational Linguistics.

Dmitry Morozov, Timur Garipov, Olga Lyashevskaya, Svetlana Savchuk, Boris Iomdin, and Anna Glazkova. 2024. Automatic morpheme segmentation for Russian: Can an algorithm replace experts? *Journal of Language and Education*, 10(4):71–84, Dec.

Antoine Nzeyimana and Andre Niyongabo Rubungo. 2022. KinyaBERT: a morphology-aware Kinyarwanda language model. // Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, P 5347–5363, Dublin, Ireland, May. Association for Computational Linguistics.

Ben Peters and Andre F. T. Martins. 2022. Beyond characters: Subword-level morpheme segmentation. // Garrett Nicolai and Eleanor Chodroff, *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, P 131–138, Seattle, Washington, July. Association for Computational Linguistics.

Marko Pranjić, Marko Robnik-Šikonja, and Senja Pollak. 2024. LLMSegm: Surface-level morphological segmentation using large language model. // Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, P 10665–10674, Torino, Italia, May. ELRA and ICCL.

Svetlana O Savchuk, Timofey Arkhangelskiy, Anastasiya A Bonch-Osmolovskaya, Ol'ga V Donina, Yuliya N Kuznetsova, Ol'ga N Lyashevskaya, Boris V Orekhov, and Mariya V Podryadchikova. 2024. Russian National Corpus 2.0: New opportunities and development prospects. *Voprosy Jazykoznanija*, (2):7–34.

Alexey Sorokin and Anastasia Kravtsova. 2018. Deep convolutional networks for supervised morpheme segmentation of Russian language. // Dmitry Ustalov, Andrey Filchenkov, Lidia Pivovarova, and Jan Žižka, *Artificial Intelligence and Natural Language*, P 3–10, Cham. Springer International Publishing.

Alexey Sorokin. 2022. Improving morpheme segmentation using bert embeddings. // Evgeny Burnaev, Dmitry I. Ignatov, Sergei Ivanov, Michael Khachay, Olessia Koltsova, Andrei Kutuzov, Sergei O. Kuznetsov, Natalia Loukachevitch, Amedeo Napoli, Alexander Panchenko, Panos M. Pardalos, Jari Saramäki, Andrey V. Savchenko, Evgenii Tsymbalov, and Elena Tutubalina, *Analysis of Images, Social Networks and Texts*, P 148–161, Cham. Springer International Publishing.

A. N. Tikhonov. 1990. *Word Formation Dictionary of the Russian language [Slovoobrazovatel'nyi slovar' russkogo yazyka]*. Russkiy yazyk, Moscow.

Dmitry Zmitrovich, Aleksandr Abramov, Andrey Kalmykov, Vitaly Kadulin, Maria Tikhonova, Ekaterina Taktasheva, Danil Astafurov, Mark Baushenko, Artem Snegirev, Tatiana Shavrina, Sergei S. Markov, Vladislav Mikhailov, and Alena Fenogenova. 2024. A family of pretrained transformer language models for Russian. // Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, P 507–524, Torino, Italia, May. ELRA and ICCL.