Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2019"

Moscow, May 29—June 1, 2019

COMPARING MODELS OF MORPHEME ANALYSIS FOR RUSSIAN WORDS BASED ON MACHINE LEARNING

Bolshakova E. I. (eibolshakova@gmail.com)
Moscow State Lomonosov University; National Research University Higher School of Economics, Moscow, Russia

Sapin A. S. (alesapin@gmail.com)
Moscow State Lomonosov University, Moscow, Russia

The paper reports on the experimental comparison of several machine learning models proposed in recent years for automatic morpheme segmentation of Russian words, including conditional random fields (CRF), sequence-to-sequence neural network (Seq2seq), convolutional neural network (CNN) model, as well as a new model we have developed with the aid of gradient boosted decision trees (GBDT). For more complete research, in our experiments we have also evaluated the semi-supervised method of Morfessor. All the morpheme analysis models being compared are briefly described in the paper, some of them perform only segmentation of words into morphs, the other produce segmentation with classification of resulted morphs. Since for Russian language linguistics rules for splitting words into morphs (and also the classification of some morphs) may differ, the experiments were performed for two data sets differing in labeling, which are obtained respectively from CrossLexica's dictionary and Tikhonov's dictionary. The experimental evaluation has shown that two best models of morpheme segmentation with classification, namely GBDT and CNN models have comparable quality, giving about 86–94% of word-level accuracy.

Keywords: morphological segmentation, morpheme analysis of Russian words, machine learning models for morphology, morpheme segmentation with classification

СРАВНЕНИЕ МОДЕЛЕЙ МОРФЕМНОГО РАЗБОРА ДЛЯ РУССКОГО ЯЗЫКА, ОСНОВАННЫХ НА МАШИННОМ ОБУЧЕНИИ

Большакова Е. И. (eibolshakova@gmail.com)
МГУ имени М. В. Ломоносова; НИУ ВШЭ, Москва, Россия

Сапин А. С. (alesapin@gmail.com)
МГУ имени М. В. Ломоносова, Москва, Россия

1. Introduction

Morpheme segmentation is the task of breaking words into constituent morphs (root and affixes), which are the smallest meaningful units of texts. This task was studied since early years of natural language processing (NLP), but not intensively, and significant progress in its solution is related with the use of machine learning techniques [4], [5], [10], [11]. Nevertheless, the task is far from complete solution, it is especially difficult for languages with rich morphologies (such as Russian) with many affixes of various types and meanings.

Nowadays, the problem of automatic morpheme segmentation became more topical, since information about morphemic structure of words is already in use in various NLP applications and tasks, including machine translation, recognition of semantically related words (cognates, paronyms, etc.), constructing word embeddings (by using morphemes as meaningful units instead of symbol N-grams), handling rare and out-of-vocabulary words (by deriving their meaning based on distributional word vectors representations) and so on.

The problem of morpheme analysis is investigated in two variants, which we consider in this paper:

- only segmentation, that is splitting a word into morphs or morpheme-like units, for example: *пре-крас-н-ый, beauti-ful*;
- segmentation with classification (categorization) of segmented morphs, by recognizing and labeling their types (the main types are Prefix, Root, Suffix, Ending), for example: *пре*:PREF/*крас*:ROOT/*н*:SUFF/*ый*:END*, beauti*:ROOT/*ful*:SUFF.

The earliest known method of morpheme segmentation was proposed by Z. Harris in [8] and based on letter variety statistics, which is counted on dictionary words. Morpheme boundaries are detected at locations in a word where the predictability of the next letter in the word is low. The method was tested with a small English dictionary (about 1000 words) and showed 61% of precision.

The last years, the most known solution for morpheme segmentation is implemented in Morfessor system [4, 5, 10], which exploits unsupervised machine learning methods to be trained on a large text collection. Evaluation of the method for English, Finnish, and Turkish showed about 70% of F-measure for detected morpheme boundaries. The method is especially useful when labeled text data needed for supervised learning are absent or insufficient. Recently, a semi-supervised method was developed, which uses some labeled data in addition to the text collection.

The task of morpheme segmentation with classification of segmented units is more relevant for morphologically rich languages (such as Russian), but it is almost unexplored, mainly because for a long time the amount of data required for supervision was not sufficient.

For Russian, the first known work on morpheme segmentation with classification through machine learning was undertaken in the morphological processor [3], whose functionality among other features provides morpheme analysis of words. The task was solved by classifying letters of words according to the main types of morphs, with conditional random fields (CRF) method. After training on the labeled data of Wiki and CrossLexica's dictionary [2], the method showed classification accuracy up to 79.5%.

In two more recent works, neural machine learning models for morpheme segmentation of Russian words were proposed. In [1], the sequence-to-sequence neural network model (Seq2seq) trained on data of Tikhonov's dictionary [12] is described. The model implements only segmentation and outperforms Morfessor's model trained on Russian text collection Librusec1.

The work [11] successfully applied convolutional neural network (CNN) for morpheme segmentation of Russian words and classification of segmented morphs. The CNN model was trained and evaluated on Tikhonov's dictionary, achieving accuracy of classification up to 88% and F-measure about 98% for morpheme boundaries, thus outperforming the previous models for Russian.

For further research of various supervised machine learning methods with respect to morpheme analysis task for Russian, we have developed one more model of segmentation with classification, applying gradient boosted decision trees (GBDT) [7] for the task2. Our choice of this method was based on the following considerations. First, compared with neural network methods, decision trees are simpler and more interpretable. They may also be applied to tasks of sequence tagging (morpheme segmentation may be treated as sequence labeling), and GBDT method is powerful enough due to boosting.

The main purpose of this paper is to describe an experimental comparison of our GBDT model and three above-mention supervised models3 for Russian on the same training data, both for morpheme segmentation task and for segmentation with classification. For this reason, we exploited two data sets with somewhat different labeling, since in some cases linguists have no consensus about how to correctly split Russian words into morphs (and also about classification of some morphemes). The first dataset contains about 23,000 segmented words taken from CrossLexica [2], the second is obtained from electronic version of Tikhonov's dictionary [12] with 90,000 segmented words.

[1] http://lib.rus.ec

[2] https://github.com/alesapin/GBDTMorphParsing

[3] n experiments we have used either open source code of them or available trained models

For both data sets, the quality of segmentation has been evaluated (with F-measure of morph boundaries), as well as classification accuracy for the models performing morpheme classification. For completeness, we have also included in the comparison semi-supervised segmentation method of Morfessor [10]. The experimental evaluation has shown that GBDT and CNN models are the best models of morpheme segmentation with classification, having comparable quality of 86–94% (depending on the datasets) measured in word-level accuracy of classification.

The paper starts with brief description of the morpheme segmentation models for Russian being compared, our new GBDT model is described in more details. Then the differences between the training data sets are explained, and results of experimental evaluation for the models are presented and discussed.

2. Morpheme Segmentation Models under Comparison

Morfessor [4], [5], [10] presents a family of statistical morpheme segmentation methods based on the maximum a posteriori estimation principle (MAP). Initially, the purely unsupervised method was developed, which works out the best variant of breaking words into morpheme-like segments for a given large unlabeled text. A semi-supervised improvement of the method was proposed later, it refines word segmentation by additional usage of yet segmented data [10]. For English and Finnish, the best results were obtained when training on text collection with 200 thou. words and additional dictionary of 10 thou. segmented words. For Morfessor 2.0, F-measure for detected morpheme boundaries increased to 77–80% for English, Finnish, and Turkish (for Russian, the method was trained but not evaluated).

The model [3] created for Russian and based on CRF method [9] performs morpheme segmentation with classification. The task was considered as sequence labeling for letters of a given word, by classifying them to four main types of morphs: Prefix, Root, Suffix, Ending. The built CRF classifier accounts for several features of a letter being classified, features of the word being segmented (its morphological tags), as well as Harris' values [8] (that are local maximums of letter frequencies counted for various positions in the words). The classification implies detecting boundaries between morphs of different classes, but cannot perform complete segmentation, since resulted segments may contain several successive morphs of the same type. For example, letters of the noun *душевность* (*soulfulness*) are classified as follows:

д	*у*	*ш*	*е*	*в*	*н*	*о*	*с*	*т*	*ь*
R	R	R	S	S	S	S	S	S	E

Successive suffixes *евн-* and *ост-* are not separated, complete segmentation should be the following: *душ*:ROOT/*евн*:SUFF/*ост*:SUFF/*ь*:END. Another weak point of such classification is inability to distinguish postfix *ся/сь* from endings. The CRF classifier trained on segmented words from Wiki and CrosLexica's dictionary has achieved 74.2% of word-level classification accuracy (percent of completely correctly analyzed words).

The sequence-to-sequence neural network model (Seq2seq) implemented in [1] for morpheme segmentation task considers this task as sequence transduction and uses ideas of encoder-decoder, originally applied for machine translation. For Russian, the model was trained on the data of Tikhonov's dictionary [12], demonstrating 93.95% of F-measure for detecting morpheme boundaries, thus outperforming by this measure both Morfessor (trained on Librusec text collection) and the considered CFR model (but Seq2seq model does not perform morpheme classification).

Significantly better quality of both segmentation and morpheme classification was achieved by convolutional neural networks (CNN) model [11], which was also trained and tested on Tikhonov's dictionary. The best reported results are 98.10% for F-measure on morpheme boundaries and 88.71% of word-level accuracy. The implemented model is quite complicated, it ensembles three CNN models with different random initializations and additionally makes use of morpheme memorizing techniques. The authors tried to add long-short memories (LSTM) layers to the network, but this did not improve the quality of the model. Compared with the above-described CRF model, it detects boundaries between successive prefixes and suffixes and also produces a more detail classification of segmented morphs, including postfix *ся/сь*, linking letter in multi-root words, and also hyphen. For this purposes, the model classifies letters of a word into 22 classes based on BMES (BIOES) labeling scheme (often used in named entities recognition task). The classes account for cases of beginning (B), middle (M), and ending (E) positions of letter within roots, suffixes, prefixes, as well as their Single letter variants. The following example shows classification for letters of the word *учитель* (*teacher*) (its segmentation is *уч*:ROOT/*и*:SUFF/*тель*:SUFF):

у	*ч*	*и*	*т*	*е*	*л*	*ь*
B-ROOT	E-ROOT	S-SUFF	B-SUFF	M-SUFF	M-SUFF	E-SUFF

It should be noted, that before evaluation the resulted morpheme classification (performed by the CNN ensemble) some segmented words are corrected by an auxiliary procedure, which fixes incorrect sequences of morph types (in particular, if a suffix is located before a root).

Our model developed for morpheme segmentation with classification is based on decision trees with gradient boosting (GBDT) [7]. Similar to CNN model, our classification of morphemes includes the main types: prefix, root, suffix, ending, and also postfix *ся/сь* (*мы-ть-ся*), linking letter for multi-root words (such as *пар-о-ход*), hyphen (*по-нашему*). In contrast to CNN model, for purposes of segmenting successive suffixes, prefixes, roots we label only beginnings of them (since the set of BMES labels is redundant for the task of morpheme segmentation). Thereby, the model classifies letters of word into 10 classes, an example for the word *учитель* (*teacher*) is given below (B-ROOT and B-SUFF encode the beginning of root and suffix, respectively):

у	*ч*	*и*	*т*	*е*	*л*	*ь*
B-ROOT	ROOT	B-SUFF	B-SUFF	SUFF	SUFF	SUFF

For classification, our GBDT model takes into account both features of the letter being classified and features of its word. The former includes the letter itself (represented in one-hot encoding format), is it a vowel, the position of the letter in the word, its occurrence frequency in training data set, and also Harris' values [8].

Since the gradient boosting method is not oriented to sequence labeling tasks, in order to account for information about sequencing of letters in our task (to be more precise, to account for influence of the previous and subsequent letters on the class of the current letter), a window of small size is used in the model: 5 letters on the left and 5 letters on the right are accounted as features (we assume that there are no more long dependencies between the letters).

Among features of the word our model includes some its morphological tags: part of speech, case, number, gender, time (if any), and stem length (we obtain all of them from the morphological parser CrossMorphy [3] with non-contextual method of morphological disambiguation).

Similar to work [11] we also have elaborated a heuristic procedure correcting some errors of GBDT classification. In particular, for word *рождаться* the segmentation obtained by GBDT: *р*:ROOT/*о*:PREF/*жд*:ROOT/*ать*:END/*ся*:POSTFIX will be corrected as *рожд*:ROOT/*ать*:END/*ся*:POSTFIX. The procedure relies on obvious rules of morphotactics for Russian: any word should begin with a prefix or root, a root may go after prefix, a suffix may go after root or another suffix, and so on.

Our GBDT model was implemented with Catboost library [6] that does not require to manually encode categorical features (such as parts of speech, etc.) into one-hot encoding. Besides implemented GBDT model, for experiments we used available implementations of the other described models4 for morpheme segmentation and classification.

3. Data Sets for Training and Evaluation

For Russian language there are two data sets with words spitted into classified morphs and thus suitable for training supervised machine learning models—an electronic version of Tikhonov's dictionary [12] with 96046 words and a subset from CrossLexica's dictionary [2] with 23426 segmented words. In both data sets, the same morpheme types (prefix, root, suffix, ending, postfix) is used, and successive prefixes and suffixes are labeled. Beside the size of the data sets, they differ significantly with respect to several important features.

First, many words represented in both data sets have different segmentation into morphs (and even the classification of some morphs), because there is no full agreement between linguists about rules of morpheme analysis. The authors of these dictionaries applied slightly different rules for splitting words into morphs and their classifying, and they also pursued different principles of forming the dictionaries. In particular, in Tikhonov's data set word *бывать* is segmented as *бы*:ROOT/*ва*:SUFF/*ть*:SUFF and in CrossLexica's set, as *бы*:ROOT/*ва*:SUFF/*ть*:END. In this case morph *ть* of verb infinitive is interpreted either as suffix or as ending, but this is unresolved question in Russian linguistics. Unlike Tikhonov's dictionary, where many prefixes are not separated from roots because of their "lexical cohesion", in CrossLexica's data all possible prefixes are usually separated even for words of foreign origin. For example, the word *продуктивный* (*productive*) in Tikhonov's data

[4] https://github.com/aalto-speech/morfessor
http:/gihub.com/alesapin/XMorphy
http:/gihub.com/kpopov/morpheme_seq2seq
https://github.com/AlexeySorokin/NeuralMorphemeSegmentation

set is segmented as *продукт*:ROOT/*ивн*:SUFF/*ый*:END, while in CrossLexica's set: *про*:PREF/*дукт*:ROOT/*ив*:SUFF/*н*:SUFF/*ый*:END (since this word has the common root with word *индуктивный—inductive*). The considered difference in morpheme segmentation are mainly explained by different functions of the dictionaries: Tikhonov's dictionary was originally built as a derivational dictionary, while CrossLexica's segmented and labeled data were created for constructing pairs of morpheme paronyms (that are words with the same root but differing in affixes and having close meanings, such as *массивный* and *массовый—massive* and *mass*).

Moreover, the considered dictionaries differs in lexicon. Tikhonov's dictionary contains many obsolete words that now hardly appear in texts, for example: *снитковый*, *лядунка*. There are no such words in CrossLexica, but it includes many words of modern lexicon, such as *эксклюзивный* (*exclusive*), *целлюлит* (*cellulite*), and so on. At the same time, in CrossLexica's data set there are no multi-root words and words with hyphen.

It is unclear a priory, which of these dictionaries is more relevant and suitable for machine learning (on our opinion, for various NLP application may be useful models trained on various data sets). For this reason, we have evaluated all above-described supervised machine learning models separately on both these data sets, except the CRF model (since only the model trained on CrossLexica's data is available). Thus, we have seven supervised models for comparison.

For evaluation, we took the available Seq2seq and CNN models pre-trained on Tikhonov's dictionary, the other models were trained and evaluated in the following way. The data for training and testing were randomly divided in proportion 80:20, the training samples were 76,836 and 18,740 words, for Tikhonov's dictionary and CrossLexica respectively. Each trained model was then evaluated with remaining samples from the same dictionary. Morfessor's semi-supervised model was trained on certain part of Librusec text collection (about 100M tokens) and was evaluated on the corresponding testing data.

4. Results of Experiments and Discussion

Our experiments with training GBDT model showed that the best evaluation scores were achieved for 10,000 iterations for the Tikhonov's dictionary and 5,000 for the CrossLexica's data, and the optimal depth of decision trees turned out to be 10.

All the machine models under comparison perform segmentation into morphs, and we have evaluated them by BPR metric (boundary precision and recall). Precision is the ratio of the number of true morpheme boundaries to the number of all revealed boundaries, while recall is the ratio of the number of true boundaries to the total number of boundaries. F-measure is computed as mean harmonic of the recall and precision. The results are given in **Table 1**.

Table 1. Comparison of Morpheme Segmentation for Russian

	CrossLexica's Dictionary			Tikhonov's dictionary		
Model	Precision	Recall	F-measure	Precision	Recall	F-measure
Morfessor	93.30	75.40	83.40	94.7	73.70	82.90
CRF	96.05	70.93	81.60	—	—	—
Seq2seq	94.62	93.92	94.27	94.07	93.83	93.95
CNN	98.68	98.75	98.72	**97.86**	**98.35**	**98.10**
GBDT	**98.84**	**99.26**	**99.05**	97.76	98.26	98.01

It turned out that CNN and GBDT models are the best for morpheme segmentation, they have very close scores and thus comparable quality. For Tikhonov's dictionary, CNN model has slightly better scores than GBDT, but the latter slightly outperforms CNN on CrossLexica's data. Seq2seq model shows quite good results, but this is the average quality. Morfessor and CRF model mainly lose in recall (because of undersegmentation) and thus in F-measure. CRF model demonstrates the worst result in F-measure (because it does not detect boundaries between successive suffixes), despite the fact that CRF method is suited for sequence labeling tasks.

In evaluation of morpheme segmentation with classification only CRF, CNN and GBDT models can participate, but we have compared only two best models. For comparison, we have evaluated classification accuracy for letters (the ratio of correctly recognized classes of letters to the number of all letters), as well as accuracy for words (the ratio of completely correctly segmented words with true classes of their segmented morphs). Table 2 presents results of the evaluation. Two evaluated models again have close scores. GBDT model insignificantly wins on CrossLexica's data, and for Tikhonov's dictionary we have the opposite situation.

However, it should be noted that while applying CNN model, 180 words from 4,686 testing words of CrossLexica's data set and 1,871 from 19,210 testing words of Tikhonov's data (that is respectively 4% and 10% of all the testing words) have been corrected by the auxiliary procedure, which fixes wrong morpheme types in the output of neural network classifier. In contrast with CNN, the analogous correction procedure of GBDT have fixed only 42 words of CrossLexica's set and 804 for Tikhonov's data (0.2% and 4% respectively). Therefore, in Table 2, we present accuracy scores obtained without the correction of the classifiers ("Uncorrected" columns in the Table). One can notice that despite GBDT is a more simple method than CNN, this method itself (without correction) works better CNN.

Table 2. Accuracy of Morpheme Segmentation with Classification

	CrossLexica's Dictionary			Tikhonov's dictionary		
		Words			Words	
Model	Letters	Corrected	Uncorrected	Letters	Corrected	Uncorrected
CNN	97.88	93.23	90.48	**96.64**	**88.71**	82.62
GBDT	**98.39**	**94.20**	**93.85**	96.40	86.54	**86.24**

Expert analysis of errors in morpheme segmentation shows that the most frequent ones are related with wrong boundaries between root and suffix, such as *печеч*:ROOT/*к*:SUFF/*а*:END instead of correct *печ*:ROOT/*ечк*:SUFF/*а*:END for the word *печечка* (*little bake*), the error occurs in both the models. Another frequent error is different segmentation of suffixes, for example: *возбуждение* (*excitation*): воз:PREF/бужд:ROOT/ени:SUFF/е:END and воз:PREF/бужд:ROOT/ен:SUFF/и:SUFF/е:END, the former is erroneous for CrossLexica's data, and the latter, for Tikhonov's data. Indeed, the problem of identificating boundaries between succsessive morphemes is especially difficult for Russian suffixes.

Table 3. Examples of wrong segmentation

Word	Model	Data	Wrong segmentation	Correct segmentation
препираться	GBDT	TN	*препир*:ROOT/*а*:SUFF/*ть*:SUFF/*ся*:PF	*препира*:ROOT/*ть*:SUFF/*ся*:PF
фанатский	GBDT	CL	фан:ROOT/ат:SUFF/ск:SUFF/ий:END	*фанат*:ROOT/*ск*:SUFF/*ий*:END
помои	CNN	TN	*помо*:ROOT/*и*:END	*по*:PREF/*мо*:ROOT/*и*:END
пришить	CNN	CL	*при*:PREF/*ши*:ROOT/*ть*:END	*при*:PREF/*ш*:ROOT/*ить*:END

Some examples of errors in complicated cases of segmentation are given in Table 3: in "Data" column, symbol TN denotes Tikhonov's data set, while CL, CrossLexica's data set. One more interesting error is the following segmentation: *по*:PREF/*душ*:ROOT/*еч*:SUFF/*н*:SUFF/*ый*:END (given by CNN on CrossLexica) instead of *под*:PREF/*уш*:ROOT/*еч*:SUFF/*н*:SUFF/*ый*:END, for word *подушечный* (*pillow's*): it seems as the model misunderstands this word as *подушевой* (*per person*).

Since trained decision trees is interpretable, we can know significance of the features have been exploited in our GBDT model (their weights, as contribution to the results). The most important features are the following: the letter being classified (10.89%), the next 2 letters (11.79% and 7.32%), 3 preceding letters (11.41%, 9.19%, and 5.74%). The Harris' values are also important, giving 4.19% for the initial part of a word and 3.12% for the end part. Morphological features of the word are less important, but give 9.22% in total.

5. Conclusions and Future Work

We have experimentally evaluated and compared five various machine learning models of morpheme segmentation for Russian, exploiting for their training two data sets with different labeling of constituent morphs. Two models of morpheme segmentation with classification, namely GBDT (the gradient boosted decision trees) and CNN (convolutional neural network) have showed the best and comparable results, and thus they may be used in various NLP experiments with Russian text. It seems that in the achieved quality of segmentation is close to possible limits for machine learning. Nevertheless, the problem of gold standard for Russian morpheme segmentation requires additional research.

As for our GBDT model, it seems reasonable to study some new features accounting for in machine learning, in order to further improve the model, supposedly by using statistics of affixes. We also plan to extend and refine the data sets for training and to combine machine learning with additional procedures based on linguistic rules.

References

1. *Arefyev N. V., Gratsianova T. Y., Popov K. P.* (2018), Morphological Segmentation with Sequence to Sequence neural network, Computational linguistics and Intellectual Technologies: Proceedings of the Int. Conference “Dialogue 2018”, Moscow, pp. 82–91.
2. *Bolshakov I. A.* (2013), CrossLexica—Universum of links between Russian words [CrossLexica—universum svyazey mezshdu russkimi slovami], Business Informatics [Biznes Informatica], No 3 (25), pp.12–19.
3. *Bolshakova E. I., Sapin A. S.* (2018), A Morphological Processor for Russian with Extended Functionality, Analysis of Images, Social Networks and Texts: 6th Int. Conference AIST 2017, Moscow, Revised Selected Papers, LNCS, 10716, Springer, Cham, pp. 22–33.
4. *Creutz M., Lagus K.* (2006), Morfessor in the Morpho Challenge, PASCAL Challenge Workshop on Unsupervised segmentation of words into morphemes, Venice, Italy.
5. *Creutz M., Lagus K.* (2007), Unsupervised models for morpheme segmentation and morphology learning, ACM Transactions on Speech and Language Processing, 4(1), Article 3.
6. *Dorogush A. V., Ershov V., Gulin A.* (2018) CatBoost: gradient boosting with categorical features support, available at: //arXiv preprint arXiv:1810.11363.
7. *Friedman J. H.* (2002), Stochastic gradient boosting, Computational Statistics & Data Analysis. Vol.38, pp. 367–378.
8. *Harris S. Zellig.* (1967), Morpheme boundaries within words: Report on a computer test, Transformations and Discourse Analysis Papers 73, pp. 68–77.
9. *Lafferty J., McCallum A., Pereira F. C. N.* (2001) Conditional random fields: Probabilistic models for segmenting and labeling sequence data, Proceedings of the 18th International Conference on Machine Learning, Williamstown, pp. 282–289.
10. *Smit P., Virpioja S., Gronroos S., Kurimo M.* (2014), Morfessor 2.0: Toolkit for statistical morphological segmentation, Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the ACL, Gothenburg, pp. 21–24.
11. *Sorokin A., Kravtsova A.* (2018) Deep Convolution Networks for Supervised Morpheme Segmentation of Russian Language, Proceedings of the Conference on Artificial Intelligence and Natural Language, St-Petersburg, Springer, Cham, pp. 3–10.
12. *Tikhonov A. N.* (1990) Word Formation Dictionary of Russian language [Slovoobrazovatel’nyi slovar’ russkogo yazyka], Moscow, Russkiy yazyk.