Lomonosov Moscow State University

as a manuscript

Sapin Aleksandr

METHODS AND TOOLS OF MORPHOLOGICAL SEGMENTATION FOR NATURAL LANGUAGE PROCESSING SYSTEMS

PhD Dissertation Summary

for the purpose of obtaining academic degree
Doctor of Philosophy in Computer Science

Moscow — 2023

The PhD dissertation was prepared at Lomonosov Moscow State University, faculty of Computational Mathematics and Cybernetics.

Academic Supervisor: Elena I. Bolshakova, Candidate of Physical and Mathematical Sciences, Docent, Lomonosov Moscow State University

1. Introduction

Topic of the dissertation and its relevance. Nowadays, the development of software systems for natural language processing (NLP) is becoming an increasingly important field in computer science. The reason for this is related to the rapid growth of digital information, primarily on the Internet.

One of the key software modules in NLP systems is morphological processors performing morphological analysis and synthesis of text *word forms*. Traditional tasks of morphological analysis include obtaining a normal form or lemma (lemmatization) from a given word form, determining its morphological tags (morphological tagging), and morphological disambiguation (resolving the ambiguity of tags). For example, for a word form *тетрадей (notebooks)*, the lemma *тетрадь (notebook)* and following morphological tags are to be recognized: noun, genitive case, plural number, feminine gender. Methods for solving these morphological analysis tasks are well studied; modern morphological processors perform them with high quality.

Besides these traditional tasks, the morphological analysis includes *morphological segmentation* task (also called *morphemic parsing*), which recognizes the internal structure of a given word by breaking (segmenting) it into morphs (morphemes), for instance: *beautiful → beauti-ful*, *прекрасный → пре-крас-н-ый*. Morphemes are the minimal meaningful units of texts and, therefore, can be taken into account in the semantic analysis of texts, which determines the relevance of this task.

To date, the quality of known developed methods for automatic morphological segmentation is insufficient for NLP applications. To solve this problem, complex linguistic features of natural languages are to be accounted for, and it is especially challenging for languages with complex morphology, for example, Russian language with a large number of suffixes, prefixes and endings.

Several approaches to automatic morphological segmentation (morphemic parsing) of words are known. A statistical approach to morphological segmentation was proposed more than a decade ago[1], however, the statistical methods had rather low accuracy. In recent years several methods based on machine learning have appeared[2]. They have improved the accuracy of morphological segmentation but consider only one aspect of the task – segmentation of normal forms of words (lemmas). However, texts consist of words in different grammatical forms (word forms), and their morphological segmentation requires additional study. Moreover, the performance aspect of software implementations for morphological segmentation methods (the rate of processed words per second and consumed memory) is of great importance in practical applications, but it has not been studied at all.

[1] Creutz M., Lagus K. Unsupervised models for morpheme segmentation and morphology learning // ACM Transactions on Speech and Language Processing. – 2007 – Vol. 1, no. 1. – P. 1–34.

[2] Ruokolainen T., et al. Painless Semi-Supervised Morphological Segmentation using Conditional Random Fields // Proceedings of the 14th EACL conference. – 2014 – P. 84–89.

Modern NLP applications and auxiliary tasks that require information about the morphemic structure of words encompass machine translation, creation of word-formation resources (derivative trees and schemes for generating new words), recognition of the meaning of new and rare words, construction of words vector representations (embeddings). While high-precision morphological segmentation methods have not been created for word forms yet, some more simple word segmentation methods are used in research works, which nevertheless improve the quality of solving downstream tasks[3]. In order to further improve the quality of solving the tasks, more accurate information about the internal structure of words is needed, which requires the development and experimental study of appropriate methods for morphological segmentation of word forms.

Since all known morphological processors for Russian language do not provide tools for morphological segmentation word forms, it is also important to create a processor that implements, in addition to the traditional morphological analysis tasks, morphological segmentation of words based on high-accuracy methods.

The goal of the dissertation work implies the development and study of morphological segmentation methods and tools that perform with high accuracy (quality), as well as performance acceptable for practical application. To achieve this goal, it is necessary to solve the following:

1. To develop and experimentally study high-accuracy (more than 88% of correctly parsed words) methods for automatic morphological segmentation of normal forms of words (lemmas) of the Russian language.
2. To develop a method for automatic morphological segmentation of Russian word forms, with accuracy not lower than methods for lemmas.
3. To realize the possibility of simultaneous morphological tagging and morphological segmentation of Russian word forms.
4. Based on the developed methods, implement corresponding software morphological processor modules which perform morphological analysis tasks with performance sufficient for practical applications (more than 10 thousand words per second on one CPU core).

2. Key results

The novelty and theoretical significance This work presents an experimental research on automatic morphological segmentation methods for Russian lemmas (normal forms) based on machine learning. Among the studied methods, the convolutional neural network method shows the best quality of morphological segmentation (89% of correctly parsed words). For Russian word forms, an automatic morphological segmentation method was first proposed. The proposed method has

[3] Hofmann V., Pierrehumbert J., Schütze H. Superbizarre Is Not Superb : Derivational Morphology Improves BERT’s Interpretation of Complex Words // Transactions of the Association for Computational Linguistics. – 2021. – Vol. 1. - C. 3594 - 1608.

shown the high quality for both morphological segmentation of word forms and morphological segmentation of lemmas (90-91%). Besides, a method for simultaneous morphological tagging of word forms and their morphological segmentation was developed for the first time, and it is implemented with high quality of both tasks. These results can be used as a basis for building morphological software models that recognize the internal structure of words, and at the same time, the results may be useful for developing morphological segmentation methods for texts in other natural languages.

Practical significance of the dissertation work is the developed open-source software library for the morphological analysis of texts in Russian, with the following features:

- It provides the tools for morphological segmentation of Russian lemmas and word forms which is useful for the implementation of NLP applications in cases when the traditional morphological and morphological segmentation are simultaneously required;
- The performance of word forms analysis, including morhological segmentation, achieves up to 20 thousand words per second on a single processor core for ongoing morphological analysis.

Main provisions to be defended.

1) A neural network method for automatic morphological segmentation of Russian word forms that worked out on the results obtained from an experimental study conducted with the models for lemmas (normal word forms). To implement the method, a procedure to automatically construct a dataset with morphologically segmented word forms was elaborated and applied. It was shown that the method for word forms is superior in accuracy to the known methods of morphological segmentation.
2) A neural network architecture was proposed as a core of the developed method for simultaneous morphological tagging word forms and their morphological segmentation, with high quality of solving both tasks. In order to increase the performance of the method, several such neural networks models were developed.
3) A software library (morphological analyzer XMorphy) implementing the developed methods and models and intended for morphological analysis and segmentation of texts in Russian with high accuracy and performance.

Personal contribution. The above-described provisions 2) and 3) were obtained solely by the author of the dissertation, and he is the main author of the papers [2; 3]. Key ideas of the developed morphological segmentation models were discussed and worked out together with the scientific adviser, E. I. Bolshakova, and the annotated dataset used for training the segmentation model for word forms was created in collaboration with the scientific adviser.

Volume and structure of the work. The dissertation consists of introduction, four chapters and conclusion. The total volume of is 89 pages including 20 figures and 15 tables. The list of references contains 84 titles.

3. Publications and approbation of the work

First-tier publications

1. Bolshakova E. I., Sapin A. S. A Morphological Processor for Russian with Extended Functionality // International Conference on Analysis of Images, Social Networks and Texts. – Lecture Notes in Computer Science, V. 10716, Springer, Cham. — 2017. — P. 22—33. — (Scopus, Q2).
2. Bolshakova E. I., Sapin A. S. Building a Combined Morphological Model for Russian Word Forms // International Conference on Analysis of Images, Social Networks and Texts. – Lecture Notes in Computer Science, V. 13217, Springer, Cham. — 2022. — P. 45—55. — (Scopus, Q2).

Second-tier publications

3. Sapin A. S. Building neural network models for morphological and morpheme analysis of texts // Proceedings of ISP RAS. — 2021. — T. 33, № 4. — C. 117—130. — (list of approved HSE journals).
4. Bolshakova E. I., Sapin A. S. Comparing models of morpheme analysis for Russian words based on machine learning // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference ”Dialogue 2019”. — 2019. — P. 104—113. — (Scopus, no quartile).
5. Bolshakova E. I., Sapin A. S. Bi-LSTM Model for Morpheme Segmentation of Russian Words // Ustalov D., Filchenkov A., Pivovarova L. (eds) Artificial Intelligence and Natural Language. AINL 2019. Communications in Computer and Information Science, V. 1119. Springer, Cham. — 2019. — P. 151—160. — (Scopus, Q3).
6. Bolshakova E. I., Sapin A. S. An Experimental Study of Neural Morpheme Segmentation Models for Russian Word Forms // Proceedings of the Computational Models in Language and Speech Workshop (CMLS 2020), CEUR Workshop Proceedings. — 2020. — Vol. 2780. — P. 79—89. — (Scopus, no quartile).
7. Bolshakova E. I., Sapin A. S. Building Dataset and Morpheme Segmentation Model for Russian Word Forms // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference ”Dialogue 2021”. — 2021. — P. 154—161. — (Scopus, no quartile).

Reports at conferences and seminars.

1. Scientific and technical seminar “New information technologies in automated systems”, MIEM NRU HSE, Moscow, Russia, April 20, 2017;
2. The 6th International Conference on Analysis of Images, Social networks and Texts (AIST 2017), Moscow, Russia, July 27-29, 2017;
3. International conference “Computational Linguistics and Intellectual Technologies: Dialogue-2019”, Moscow, Russia, May 29 - June 1, 2019;
4. International conference “Artificial Intelligence and Natural Language. AINL 2019”, Tartu, Estonia, November 20-22, 2019;

5. Conference Lomonosov Readings 2020. Section of Computational Mathematics and Cybernetics, Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, Online, October 21 - November 2, 2020;
6. XVI TEL International conference on computational and cognitive linguistics, Online, November 12-13, 2020;
7. International conference “Computational Linguistics and Intellectual Technologies: Dialogue-2021”, Online, June 16-19, 2021;
8. The 10th International Conference on Analysis of Images, Social Networks and Texts (AIST 2021), Tbilisi, Georgia, December 16-18, 2021;
9. Scientific seminar of the Department of Intelligent Information Technologies, Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, December 23, 2021.

4. Content of the work

The **introduction** describes the research area, shows the work’s relevance, reveals its goals and objectives, and describes the novelty and practical significance of the work.

The first chapter overviews the existing morphological analysis methods for modern natural language processing (NLP) systems.

Section 1.1 introduces the basic concepts, terms and problems related to automatic morphological analysis and synthesis of texts in natural language and also describes metrics to evaluate quality of morphological analysis tasks.

Section 1.2 considers methods of morphological analysis based on dictionaries of stems and dictionaries of word forms. Such methods make it possible to solve the problems of lemmatization and morphological tagging but require additional tools for morphological disambiguation and heuristic rules for processing out of vocabulary words.

Section 1.3 describes main methods for morphological disambiguation (resolving morphological ambiguity) using dictionaries, statistical information, and machine learning methods for dictionary-based systems.

Section 1.4 considers an approach to morphological analysis based on machine learning and vector representation of words (embeddings), which allows to get rid of the morphological dictionary and achieve the best quality for lemmatization, morphological tagging and disambiguation: for the Russian language up to 96.5% of correct lemmas and 95% of correct morphological tags[4]. However, the methods of this approach significantly depend on the training data, and their software implementations have low performance.

[4] Lyashevskaya O. N., et al. GRAMEVAL 2020 Shared Task: Russian Full Morphology and Universal Dependencies Parsin // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference ”Dialogue 2020”. – 2020

Section 1.5 presents an overview of approaches to the problem of morphological segmentation (morphemic parsing) of words. Two variants of the task are considered: *morphemic segmentation*, i.e., splitting a word into its constituent morphs (for example, *зануда (bore)* → $\underbrace{\textit{за}}-\underbrace{\textit{нуд}}-\underbrace{\textit{а}}$) and morphemic segmentation with classification, when, in addition to splitting a given word, the types of resulting morphs is recognized (for example, *зануда (bore)* → $\underbrace{\textit{за}}_{\text{prefix}}-\underbrace{\textit{нуд}}_{\text{root}}-\underbrace{\textit{а}}_{\text{ending}}$).

For the problem of morphemic segmentation, the known methods are considered, based on statistics and unsupervised learning, as well as methods based on supervised learning, the latter has appeared only in recent years. The best quality for the Russian language is achieved by the supervised method based on convolutional neural networks[5] (88% of correctly segmented words); however, this method allows to segment only lemmas (normal forms) and therefore is not suitable for processing texts, which consist of word forms.

In general, the methods of morphological segmentation of lemmas have not been sufficiently studied, and for texts word forms they have not been developed at all. Thus, a additional study of the morphological segmentation methods for lemmas is needed, as well as the development of a method intended to processing word forms.

Section 1.6 compares the functions (features) and the performance (in words per second and memory consumption) of freely available morphological processors for the Russian language. It is shown that these processors implement only a part of the functions of morphological analysis and synthesis, and the function of morphological segmentation is absent. Therefore, it is relevant to develop a morphological processor with extended morphological functionality (traditional moprhological analysis and morphological segmentation) and high quality and performance.

The second chapter describes methods of morphological segmentation for Russian lemmas developed with machine learning. The more complex variant of moprhological segmentation, i.e. segmentation words into morphs with the classification of their types is investigated.

Section 2.1 describes features of labeled datasets with segmented morphs of Russian lemmas: RuMorps-Lemmas[6] (96 thousand lemmas) and RuMorphs-CrossLexica (27 thousand lemmas). The markup includes seven types of morphs: PREF (prefix), ROOT (root), SUFF (suffix), END (end), POSTFIX (postfix, *ся* and *сь* of verbs), HYPH (hyphen), LINK (vowel connecting parts of complex words), e.g., *бобриха* (*beaver*) – *бобр:ROOT/их:SUFF/а:END*.

Section 2.2 shows that the problem of morphemic segmentation with classification can be solved as a sequence labeling for letters of words along with classification of morphemes types. Depending on the set of classes, either the task of morphemic segmentation with the classification of groups of morphs of the same type

[5] Sorokin A., Kravtsova A. Deep convolutional networks for supervised morpheme segmentation of Russian language // Conference on Artificial Intelligence and Natural Language. – CCIS, Springer, Cham. – 2018

[6] https://cmc-msu-ai.github.io/NLPDatasets/

is solved (seven classes, successive morphs of the same type are not separated from each other), or the task of segmentation with the classification of morphs (ten classes, successive consecutive morphs of the same type separated from each other). In order to reveal the best morphological segmentation methods for lemmas, the following machine learning methods were choosen:

- conditional random fields (CRF);
- gradient-boosted decision trees (GBDT);
- recurrent neural network based on long short-term memory (LSTM);
- one-dimensional convolutional neural network (CNN).

Training each particular machine learning method on the labeled dataset yield software model of morphological segmentation.

At the end of the section, metrics for the quality evaluation of morphological segmentation are described. *Precision*, *Recall* and *F1-measure* are used to evaluate recognized morpheme boundaries:

$$Precision = \frac{TP}{TP+FP};\ Recall = \frac{TP}{TP+FN};\ F1 = \frac{2TP}{2TP+FP+FN} \quad (1)$$

where TP is the number of correctly recognized boundaries between morphs, FP is the number of falsely recognized boundaries, FN is the number of unrecognized boundaries. The $Accuracy$ metric is used to evaluate the correctness of the classification of letters in segmented words:

$$Accuracy_{letters} = \frac{\sum_{i=0}^{len(dataset)} \sum_{j=0}^{len(word_i)} correct(letter_j)}{\sum_{i=0}^{len(dataset)} len(word_i)}, \quad (2)$$

where $len(dataset)$ is the number of words in the considered dataset, $word_i$ is $i-$th word in the dataset, $len(word_i)$ is the length of $i-$th word, $correct(letter_j) = 1$ only when the class of the letter is correct, and 0 otherwise. Correctness of the classification of all segmented morphs for all words in a the given dataset is also evaluated as follows (i.e. accuracy of classification by words):

$$Accuracy_{words} = \frac{\sum_{i=0}^{len(dataset)} correct(word_i)}{len(dataset)}, \quad (3)$$

where $len(dataset)$ is the number of words in the dataset, $word_i$ is $i-$th word in the dataset, $len(word_i)$ is the length of $i-$th word, $correct(word_i) = 1$ only when the types and boundaries of all word morphs are recognized correctly, and 0 otherwise. This metric is the main to evaluate quality of morphemic segmentation with classification, because it takes into account all the previous metrics.

The performance of software models for morphological segmentation is evaluated as the number of processed words per second on one processor core of the Intel i7-10850H CPU (on a text collection fragment with volume of 10 million words[7]). The amount of consumed memory in megabytes (MB) is also estimates.

Section 2.3 describes how the conditional random fields (CRF) method was applied to morphemic segmentation with classification of groups of morphs (classification of letters into seven classes).

The following features are used for training: the letter itself, is it vowel or not, morphological features of the word to be segmented (part of speech, case, etc.). For training and validation the RuMorphs-CrossLexica dataset is used, evaluation of the resulted model show only 74.2% accuracy of the classification of words.

Section 2.4 describes the application of Gradient Boosted Decision Trees (GBDT) to morpheme segmentation with classification, where adjacent morphs of the same type are to be separated from each other. To solve the task, three classes have been introduced for the initial letters of prefixes, roots, and suffixes; for example, for word *торговец* (merchant) → *торг*:ROOT/*ов*:SUFF/*ец*:SUFF) the result classification is:

т	о	р	г	о	в	е	ц
B-ROOT	M-ROOT	M-ROOT	M-ROOT	B-SUFF	M-SUFF	B-SUFF	M-SUFF

The GBDT method is not a sequence labeling method, so fixed-size text window is used as features to train: 5 letters to the left and 5 to the right for the processed letter. Other features are the same as for the CRF-based model. RuMorphs-Lemmas and RuMorphs-CrossLexica datasets are used for training and validation, giving two software models.

Experiments with the GBDT model show the high quality of morphological segmentation for the RuMorphs-Lemmas dataset (86.5% accuracy for words) and the best quality for RuMorphs-CrossLexica (94.2% accuracy for words). The GBDT model made it possible to measure the significance of the features taken into account for training: the greatest influence on letter class recognition achieve by neighboring letters (the previous and two subsequent ones), as well as part of speech (POS) of the word.

Section 2.5 describes the proposed long short-term memory (LSTM) neural network architecture for morpheme segmentation with classification, as well as its training and evaluation. Since the class of a letter is mostly affected by the successive and previous letters, a bidirectional LSTM network (BiLSTM) was chosen. It was experimentally found that the best quality is achieved by a multilayer network (three BiLSTM-layers) with dropout layers between them, and the final fully connected (dense) neural network layer (Figure 1).

To train the network, the same features of the letters and the word are used as in the GBDT-based method. An ensemble of three similar BiLSTM models shows a quality of 89.03% accuracy for words when trained on the RuMorphs-Lemmas dataset and 94.5% when trained on the RuMorphs-CrossLexica dataset.

Section 2.6 describes the architecture of a convolutional neural network (CNN) for morphological segmentation task, it is based on several one-dimensional

[7] `librusec.pro` (fragment at the link `https://bit.ly/3typZ57`)

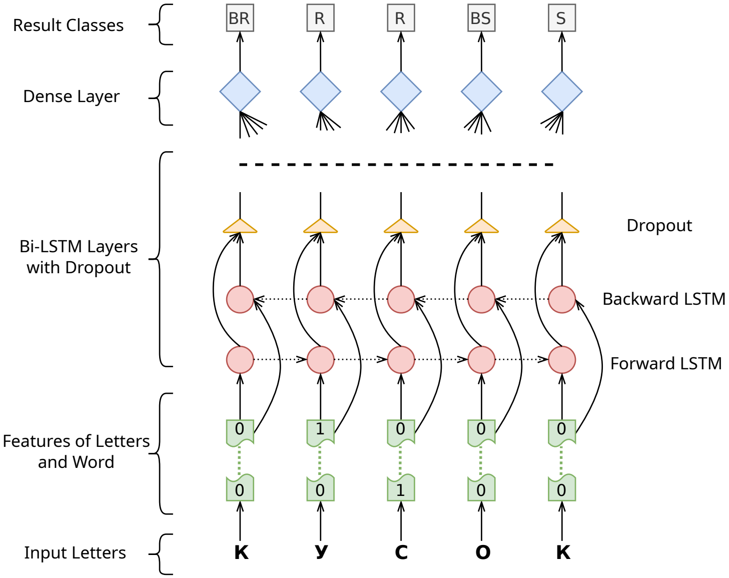

Figure 1 — Architecture of the BiLSTM morphological segmentation model

convolutional networks with dropout between them and the final dense layer for classification. The network input receives words of 20 letters, shorter words are supplemented with whitespace (insignificant) characters, and longer ones are divided into parts.

To train the network the same features were used as in the GBDT-based method. For the RuMorphs-Lemmas dataset, the trained model shows 89.5% of accuracy for words, and for the RuMorphs-CrossLexica dataset, 94.7%. These results are the best for morphological segmentation of Russian lemmas among the considered and evaluated models.

Section 2.7 compares the developed morphological segmentation software models in terms of accuracy for words ($Accuracy_{words}$ – Table 1), and in terms of performance (Table 2). The CNN-model shows the best word classification quality and the best performance, outperforming the previously proposed convolutional model[8] (with 88.6% accuracy of word accuracy).

The methods of morphological segmentation described in this section are intended for the segmentation of Russian lemmas (normal forms). However, texts consist of significantly varying word forms and it is necessary to segment not lemmas but various word forms. An experimental evaluation of the quality of morphological segmentation for word forms using the best CNN model showed less than 48% of

[8] Sorokin A., Kravtsova A. Deep convolutional networks for supervised morpheme segmentation of Russian language // Conference on Artificial Intelligence and Natural Language. – CCIS, Springer, Cham. – 2018

Table 1 — Accuracy of morphological segmentation methods of lemmas

Model	**RuMorphs-Lemmas**	**RuMorphs-CrossLexica**
CRF	-	74.2
GBDT	86.54	94.20
BiLSTM	89.03	94.49
CNN	**89.51**	**94.72**

Table 2 — Performance of software models of morphological segmentation of lemmas

Model	Words per Second	Size of Model (MB)
CRF	47	17
GBDT	269	2651
BiLSTM	64	203
CNN	**673**	**4.7**

the accuracy for word classification. The reason relates to morphologically complex Russian language, there are usually significant difference in morphemic structure of various word forms for a particular lemma, for example:

segmentation of lemma:	*расшить – рас:PREF/ши:ROOT/ть:END*
segmentation of wordform:	*разошьют – разо:PREF/шь:ROOT/ют:END*
segmentation of lemma:	*лечь – ле:ROOT/чь:END*
segmentation of wordform:	*ляжет – ляж:ROOT/ет:END*

Thus, for practical application a method for morphological segmentation word forms is required.

Chapter 3 describes the developed methods of morphological segmentation for Russian word forms, as well as the labeled datasets (previously absent) created for this purpose.

Section 3.1 describes an automatic procedure for building a dataset necessary for training a morphological segmentation model for word forms. The procedure receives lemmas and their segmentations from the RuMorphs-Lemmas dataset as input. For each lemma, using morphological dictionaries, the procedure generates word forms and determines their part of speech (POS). To segment all generated word forms, the procedure uses their part of speech, as well as grammatical information about suffixes and endings in Russian language.

The built dataset RuMorphs-Words[9] contains 2.8 million word forms with morphemic labels, including 28% of nouns, 45% of adjectives and participles, 27% of verbs and 0.05% of adverbs.

In **Section 3.2** the morphological segmentation method of Russian word forms is considered. Since the best quality and performance were shown by the CNN model, a similar neural network architecture was taken as the basis for a method for word forms.

[9] https://cmc-msu-ai.github.io/NLPDatasets/

To train the neural network, the same features of letters (the letter itself, is it vowel) and the part of speech of the word form to be segmented from the built dataset are used.

The architecture of the developed software model (see Figure 2) is based on “convolutional blocks”, consisting of a one-dimensional convolutional layer, a subsampling layer (*max pooling*), and a dropout layer. The max pooling layer allows to significantly speed up both training and inference of the model, and the dropout layer helps to avoid overfitting. In total, the model uses three sequentially-connected “convolution blocks”, the output of the last block is fed to the input of dense layers of the network (each letter of the word has its own dense layer of the network).

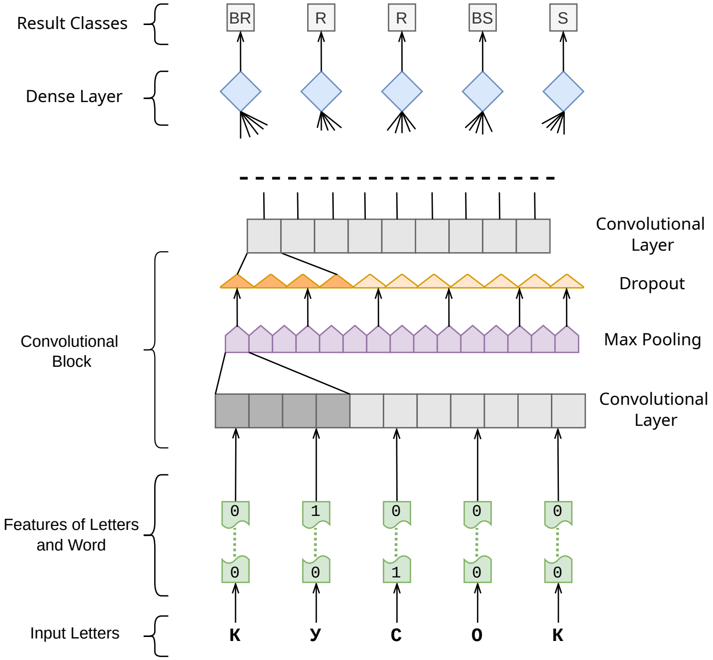

Figure 2 — Architecture of CNN-model for morphological segmentation of word forms

The CNN model trained on the built RuMorphs-Words dataset for word forms shows 91.06% of accuracy for words, while on the lemmas the quality is also high, 90.03% accuracy for words. Performance of the model is 4559 words per second without taking into account the time spent to refine POS tag of the segmented word and 2380 words per second taking part of speech tagging into account. The stage of POS tagging makes the model not only less performant, but also less convenient to use, so a method was developed that simultaneously performs a morphological analysis of word forms (including tagging) and also their morphological segmentation.

Section 3.3 proposes an architecture for a combined model of morphological analysis and segmentation of word forms.

Similar to the developed CNN model of morphological segmentation of word forms, the architecture is based on convolutional neural networks, namely convolutional blocks. Unlike the model for word forms, the combined model processes the input text by sentences (sequences of 9 words).

The architecture of the combined model (Figure 3) includes a submodel responsible for morphological disambiguation (on the left), as well as a submodel responsible for morphological segmentation (on the right).

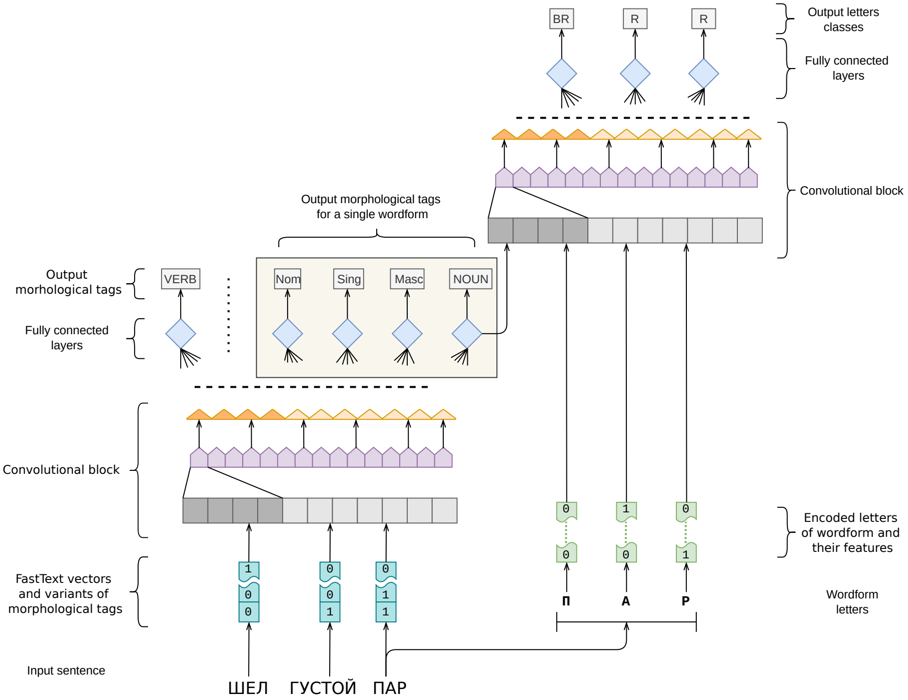

Figure 3 — Architecture of combined morphological model

Along with each word form, input of the combined model include its non disambiguated morphological tags (which are taken from the morphological processor[10]). The combined model disambiguate the tags (refining the part of speech, case, number, gender, tense), and then the disambiguated part of speech is used to perform morphological segmentation.

To train a combined model a labeled dataset is needed, contain both morphological and morphemic labels of word forms. Such datasets did not exist and in order to create appropriate one, the annotated corpus with morphological tags

[10] https://github.com/alesapin/XMorphy

SynTagRus[11] was additionally labeled with morphemic labels for each word form, giving the resulted dataset RuMorphs-SynTagRus[12].

For training a combined morphological model, the following features are used: all possible variants of morphological tags of input word forms, embeddings of word forms, and encoded letters of the word forms.

Evaluation of the combined model trained on the obtained RuMorphs-SynTagRus dataset showed *overfitting*. To solve the problem further training was divided into three stages using the *transfer learning* technique. At the first stage, the morphological segmentation submodel is trained on the dataset RuMorps-Words with word forms. At the second stage, the obtained weights of the morphological segmentation submodel are frozen (i.e. excluded from training), and the combined model is trained on the RuMorphs-SynTagRus dataset. At the third stage, the learning rate is reduced by two orders of magnitude and all combined model is trained on the RuMorphs-SynTagRus dataset.

As a result, so trained combined morphological model shows a high quality of disambiguation: 94.2% of correct morphological tags. The quality of morphological segmentation is the best for the RuMorphs-Words dataset (91.7% morphological segmentation accuracy for words) and at the same time quite high for the RuMorphs-SynTagRus dataset (88.6% accuracy for words).

The performance of the combined model implemented with using the tensorflow-lite library turned out to be 1893 words per second, which is comparable to the model for morphological segmentation of word forms.

Section 3.4 suggests a way to apply the developed models (*inference*) to word forms. Since input of neural networks (including convolutional networks) has a fixed size, the input data (words and sentences of varying size) is often padded with placeholders to a fixed size required for the model: up to 20 letters for the morphological segmentation model and up to 9 words for combined model. At the same time, for real Russian texts, the majority of words contain less than 20 letters, and there are often sentences shorter than 9 words.

To improve performance of model, we proposed to use a complex of several models with different input lengths: 5, 7, 9, 12 and 15 letters for the CNN model of morphological segmentation, and for the combined model – 5, 7, 9 words in sentences and 6, 12 and 20 letters in words respectively. For each input sequence, depending on its length, the most suitable model with the smallest input size is selected.

Using the complex of models significantly speeds up text processing (Table 3), although it increases the overall size of software model.

Chapter Four is devoted library implementation of the open-source morphological processor XMorphy for Russian language[13].

Section 4.1 gives an overall description of the processor, its functions and structure.

[11] https://universaldependencies.org/treebanks/ru_syntagrus/index.html

[12] https://cmc-msu-ai.github.io/NLPDatasets/

[13] https://github.com/alesapin/XMorphy

Table 3 — Performance of morphological segmentation models of word forms

Model	Words per Second	Size of Model (MB)
CNN	4559	1.1
Combined	1893	2.8
CNN (compelx)	7512	5.4
Combined (complex)	3543	33.5

XMorphy is based on the OpenCorpora[14] word form dictionary with the conversion of the morphological tags of this dictionary into the Universal Dependencies[15] format (which is becoming the in fact the standard for creating annoted text corpora). XMorphy processor supports the following features:

- tokenization;
- lemmatization and morphological tagging;
- morphological synthesis;
- morphological disambiguation;
- morphological segmentation.

The XMorphy processor is implemented as a C++ library and a set of command line utilities (CLI, dictionary building utilities). The source code is divided into logical modules according to the main implemented functions (Figure 4).

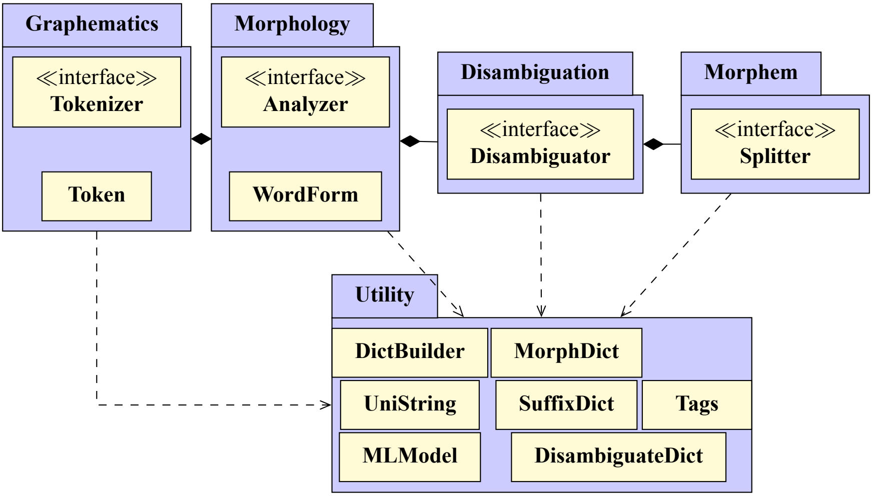

Figure 4 — Diagram of XMorphy processor modules

[14] http://opencorpora.org/
[15] http://universaldependencies.org/

Section 4.2 describes the dictionaries used in XMoprhy and the data structures for them. To store more than 5 million word forms of the morphological dictionary, a directed acyclic word graph is used, DAWG [16], the keys in the DAWG are word forms. To reduce the size of the memory consumed by the dictionary, morphological tags are grouped according to inflectional classes of Russian language word forms, so only internal numbers of classes and word endings in these classes are presented in DAWG, and the tags themselves are stored in a separate array (Figure 5).

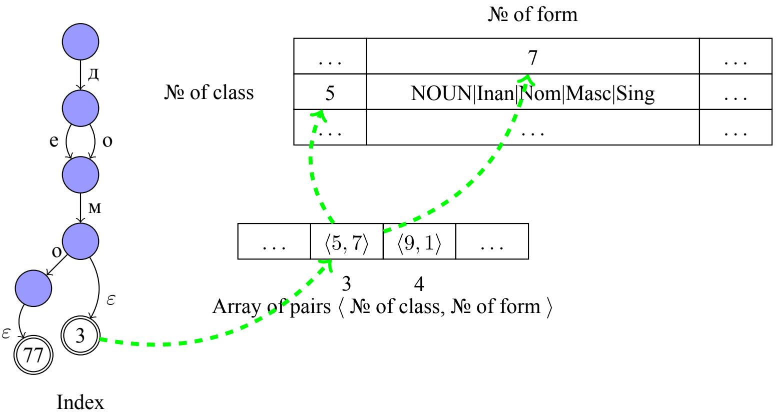

Figure 5 — Structure of morphological dictionary

The morphological analysis of a given word form is performed as its lookup in the DAWG dictionary, returning all possible variants of morphological tags. Lemmatization is performed as cutting off the ending of the given word form and concatenating ending of lemma (the normal form from the dictionary). The synthesis of a needed word forms is implemented in a similar way.

To process a word form that is absent in the morphological dictionary, cutting off known prefixes from it and then searching the remaining part in the DAWG dictionary is performed. For unknown words the analogy principle is also used to find morphological tags and lemma, for example, for unknown word *кринжовая (cringe)*, the analogy with the ending *овая* is used:

$$\textit{кринж}\underbrace{\textit{овая}} \rightarrow \begin{array}{c}\text{овая (}\textit{дворовая, бредовая, ...}\text{)}\\ \text{ADJ|Nom|Pos|Fem|Sing}\end{array} \rightarrow \begin{array}{c}\text{кринжовая} \rightsquigarrow \text{кринжовый}\\ \text{ADJ|Nom|Pos|Fem|Sing}\end{array}$$

Section 4.3 describes two implemented methods for morphological disambiguation. The statistical (contextless) method uses the statistics of occurrences of each morphological tags combination, calculated with the annotated SynTagRus corpus.

[16] Daciuk J. [et al.]. Incremental construction of minimal acyclic finite-state automata // Computational linguistics. – 2000. – Vol. 26, no. 1. – P. 3-16.

For contextual disambiguation convolutional neural networks is used, its software model has architecture similar to the morphological submodel of the combined model and it is used in cases where it is necessary to perform only disambiguation word form without morphological segmentation.

Section 4.4 describes the implementation of the morphological segmentation model for individual word forms and also implementation of combined morphological model for text analysis.

XMorphy processor has a built-in dictionary of segmented word forms, stored in the DAWG structure, containing all the word forms of the RuMorphs-Words dataset: the key is the word form and part of speech, and the corresponding value is the segmented word form. If the word form been processed is absent in the dictionary, the developed complex of CNN-models for morphological segmentation is used. A cache of previously parsed words with an algorithm for replacing least recently used elements (LRU) gives significant performance improvement of morphological segmentation of word forms.

For processing a text with word forms the complex of combined morphological models is exploited in XMorphy. The input for the selected model are variants of the morphological tags of the given word forms of the processed sentence (obtained from the processor's dictionary or predicted for non-dictionary words), their FastText embeddings[17], and also letters of word forms.

Section 4.5 characterizes the technical details of XMorphy implementation. The dynamic library that implements the processor depends only on the system standard libraries and can be used in any Linux environment. All dictionaries and models are built into the library, so it can be used without additional configuration, while the ability to dynamically use custom dictionaries and models is also available.

Due to the architecture of convolutional neural networks, the caching meachanism, the efficient library implementation of neural networks with tensorflow-lite, and the optimal compiler flags, it is possible to achieve performance of up to 20 thousand words per second for determining morphological tags, lemma and morphological segmentation on a single processor core of the Intel i7-10850H CPU.

In conclusion the main results of the work are the following:

1. Four machine learning methods for automatic for morphological segmentation of normal forms of Russian lemmas (normal forms of words) were developed and evaluated. Based on the results of their experimental evaluation, the model with the best quality segmentation for lemmas was choosen for further research.
2. The procedure for automatic building of dataset with segmented Russian word forms has been created and the dataset RuMorphs-Words was built.
3. Using the built dataset, a neural network method for morphological segmentation of word forms has been developed, showing high segmentation quality and performance.

[17] Bojanowski P., et al. Enriching word vectors with subword information // Transactions of the Association for Computational Linguistics. – 2017. – Vol. 5. – P. 135-146.

4. A neural network architecture was proposed for simultaneous morphological tagging of text word forms and their morphological segmentation and corresponding software model implemented, with a high quality of solving both tasks.
5. Using the developed models, a software library (morphological processor XMorphy) was implemented for lemmatization, morphological tagging, morphological disambiguation, as well as morphological segmentation of Russian word forms.