Automatic Morphemic Analysis of Russian Words

Lyudmila Maltina[0000-0002-3789-9035] and Alexey Malafeev[0000-0002-8962-7496]

National Research University Higher School of Economics
Nizhny Novgorod, Russia
lpmaltina@gmail.com; aumalafeev@hse.ru

Abstract. The paper considers the task of the morphemic analysis of Russian words and compares the efficiency of several proposed models. These models can be divided into three groups: derivational and inflectional rule-based, probabilistic, and hybrid models. The latter achieved state-of-the-art results of 0.848 F-score on a test set of 500 Russian words. The models use dictionaries of morphs and words and information about the part of speech and other morphological features of the word. Importantly, our solution takes into account synchronic word-formative relations between words. This allows for analyzing words in any grammatical form, as well as previously unseen words. Our system, which we make freely available to the community, also features morphemic annotation of entire texts and search for specified morphs.

Keywords: natural language processing, morphemic analysis for Russian, derivational and inflectional rules, probabilistic models, morphemic annotation of texts, search of morphs

1 Introduction

Systems that perform automatic morphemic analysis have a wide scope of application. They can be used in machine translation to reduce the volume of dictionaries and to recognize multi-morpheme out-of-vocabulary words [1-3]. Automatic morphemic analysis can be applied in morpheme notation [4] and speech recognition [5-7]. Also, it can be used for morphemic and derivational annotation of corpora, which allows one to study the functioning of the word-formation system, including the emergence ofneologisms and occasionalisms. This idea was put to use in the annotation of the Russian National Corpus [8], but a morphemic annotation / search tool can be useful for analyzing any other corpus in Russian. Other areas of application include search engines in which query expansion can be performed by finding words with the same root as the query terms [9], as well as checking and self-checking morphemic analysis performed by students [10-12].

In our study, we considered the existing approaches to automatic morphemic analysis and developed our own models based on a set of rules and on a probabilistic model. Our morphemic analysis tool [23], implemented in Python 3, also features functions for text annotation and morph search. Our system uses a morph database derived from the morpheme and spelling dictionary by A.N. Tikhonov [13] and the 1980 Russian grammar [14], as well as morph frequency and position data necessary for creating a probabilistic model. This data, as well as gold standard word analyses used for testing, was extracted from the same sources. We compared the performance of our system with one of the very few available tools [15]. This choice was made because this system, unlike [10-12], is able to analyze previously unseen words and words that are in non-initial forms. Unfortunately, it is quite difficult to find a system for the morphemic analysis of Russian words in the public domain.

2 Literature Review

Approaches to automatic morpheme analysis differ depending on the application and the databases used. The latter can be morph-only databases, or they can be supplemented by databases of words or stems. Moreover, it is possible to automatically obtain a morph list from a corpus or a word list.

A.A. Karpov [6, 16] uses morphemic analysis for speech recognition, relying on both morphological and morphemic dictionaries. If the word is not found in the morphological dictionary, then it is assumed that all of it is the root. If the word has been found, the 'ending' of the word is marked (the part consisting of suffixes, inflections and postfixes), and then the prefix is cut off.

P.V. Dikiy and M.V. Edush [17] do not use large word or stem databases to increase the processing speed of their solution; they rely on morpheme dictionaries only. Their system searches for prefixes and inflections, and then roots and suffixes by iterating over the characters of the word. With a complete match, the morph is marked in the word, while with a partial match, the search for morphs of this type continues, and if there are no matches, the search for morphs of the next type is carried out.

S.G. Fadeev and P.V. Zheltov [18] use only morph dictionaries for morphemic segmentation of words. To optimize the work of the program, morph arrays are created; these are sorted according to morph frequency and the position of the morph in the group of morphs of this type. Because of ambiguity, it is necessary to continue matching morphs even after finding the first match, but here, too, the number of checks can be reduced. If the morphological features of the language prevent the appearance of a given morph in some position (for example, one word cannot have two verbal endings), it is necessary to start looking for morphs of the next type. If some morph in some position does not occur in the database, but it is unknown whether in principle it can appear in this position, before proceeding to the search for morphs of the next type, *D* elements must be checked. The adjustable parameter *D* was called by the authors the depth of morphemic analysis.

M.G. Tagabileva and Yu.N. Berezutskaya [8, 19] used both affix and word databases. The researchers applied morphemic analysis for the annotation of the Russian

National Corpus and implementing morpheme search functionality. Search in the main subcorpus features the "derivation" option, which allows to take into account alternations within the same morpheme. When searching for a morpheme, the user can specify its position and type (prefix, root, suffix or inflection) [19]. As far as the morphemic analysis model is concerned, first prefixes and suffixes are marked, then roots are cut off. To mark prefixes, the authors take advantage of the fact that most words with prefixes have unprefixed pairs. An algorithm for extracting prefixes in words with related roots was developed. It is possible to obtain several morphemic analyses of the word, of which the correct variants is selected manually [8].

O.V. Kukushkina [9] tackles the task of finding related words and uses both affix and stem dictionaries to disambiguate word roots. The author's system is based on the principle that prefixes and roots are marked in a strictly morphemic fashion, while affixes are cut off formally, with no regard of their true morphemic boundaries. This is due to the fact that the boundary between the prefix and the root is important for finding the correct root, while finding boundaries within a suffix group does not affect root boundaries.

D. Bernhard [20] also solves the problem of finding related words. The authoruses unsupervised learning, that is, unannotated data is entered as input: a list of words without morphemic boundaries or morph types indicated. This work is based on a combination of three methods: considering the predictability of a word part, word comparison and optimization. When detecting the most predictable word parts, transition probabilities are applied. Word comparison is used to find sub-strings discriminating words. Optimization consists in the use of the length and frequency of the morph to select the right morphemic analysis.

A.S. Sapin and E.I. Bolshakova [21] developed the morphological analyzer CrossMorphy, one of the functions of which is automatic morphemic decomposition. Considering this problem as a classification problem within the framework of machine learning, the authors apply Conditional Random Fields. As the training set, ready-made morphemic analyses were taken from the dictionaries of the CrossLexis system (23,400 words) and the Wiktionary (94,400 words). The accuracy for these resources was 0.79 and 0.69, respectively.

As can be seen, the existing systems for morphemic analysis of Russian words use a variety of methods and approaches. In many cases, unfortunately, accuracy is not reported. We decided to develop our own system, with a few distinctive features in mind. In particular, our system takes into account the derivational connections between words, the part of the speech of the analyzed word and its morphological features. Additionally, the system is able to process word forms that are different from the lemma. The program also allows for accurately analyzing out-of-vocabulary and complex words, as well as perform morphemic annotation of arbitrary texts.

3 Developing a System for Morphemic Analysis

3.1 Basic Rule-Based Model (*rules*)

Lists of prefixes, suffixes, postfixes and repeated elements of complex word formation patterns were derived from the Russian grammar [14]. The morphs of the latter group were designated as "recurring elements". The suffix and postfix lists were distributed in accordance with the parts of speech they are characteristic of.

First, postfixes, inflections and form-building suffixes are found, then prefixes and suffixes are marked in words formed by the prefixed-suffix method. If prefixes have not been marked by this stage, then prefixes and "recurring elements" are selected. Then, depending on the part of speech of the analyzed word, other suffixes are found.

Postfixes, inflections and form-building suffixes are successively cut off from the word end. If the end parts of several morphs coincide, then the longest morph from this group is chosen. First, the part of the speech of the word is determined. Postfixes and form-building suffixes are marked for static parts of speech: the infinitive (INFN), verbal participles (GRND), the comparative (COMP), adverbs (ADVB) and stativewords (PRED). Next, the inflections and form-building suffixes of the morphologically-rich parts of speech are cut off. In the declinable parts of speech: noun-like pronouns (NPROs), full adjectives (ADJF), numerals (NUMR), full participles (PRTF), nouns (NOUN) - the marking of the inflection is based on letter-by-letter comparison of word forms of different numbers and cases. For unchangeable words, the inflection is not marked, but some other words may have the zero inflection. When comparing symbols, the model takes into account the possible alternation of sounds. Then form-building suffixes are marked for full participles, adjectives and comparative degree adverbs. The next are the inflections of finite verbs (VERB). For verbs in the indicative mood of the present and future tense, the inflection is marked by means of conjugating the verbs. In order to find the inflection of past tense verbs and conditional mood forms, these words are inflected for gender and number. Then, for verbs in the imperative mood, inflections and form-building suffixes are marked. After that, the inflections and suffixes of short adjectives (ADJS) and short participles (PRTS) are found.

At the next stage, words formed with the prefix-suffix method are considered. For these, possible formants (derivational affixes) and base words are established. If the hypothetic base word is found in the dictionary, then the prefix and suffix are marked in the analyzed word. For example, in the word *собеседник* (conversation partner), the prefix *со-* and the suffix *–ник-* are properly found, because there is the base word *беседа* (conversation) in the dictionary.

The next step is the marking of prefixes. If an unprefixed pair is found for the word, then the corresponding prefix is marked. For example, in the word *принесём* ([we] will bring) the prefix *при-* will be marked, because for the lemma *нести* (to bring) the unprefixed pair *нести* is found. If the beginning of the analyzed wordform coincides with the prefix, but the corresponding unprefixed pair is not found, then it is possible that the word contains a bound base (for example, as in the word *поднять* – to lift). Therefore, the system checks whether there are words in the dictionary, the first part of which is a prefix, and the remaining part coincides with the unprefixed part of the analyzed word (for the example word given above such a word is *при-нять* – to accept, which has the same root as *под-нять*). If at least one such word is found, then the corresponding prefix is marked. Similarly to prefixes, "recurring elements" are found.

Next, word-building suffixes are marked. Taking into account the derivational connections between words makes it possible to correctly identify suffixes, even if character sequences are the same. The rest of the word is considered the root.

3.2 Improved Rule-Based Model (*rules_corrected*)

This model is a modification of the previous one. By removing prefixes, suffixes and inflections from the list of all morphs extracted from the dictionary of A.N. Tikhonov [13], a list of roots was obtained. To prevent excessive marking of prefixes, the following conditions were set: the prefix is not marked if the word is less than three characters in length or the word starts with an element that is found in the list of roots and the root is one character longer than the prefix that was originally found in the word.

3.3 Maximum Matching Model (*maxmatch*)

For this and other probabilistic models below we used 100 614 lemmata from the dictionary, which comprise 17 017 different morphs. In this model, a part of the word is considered a morph if it is included in the list of morphs and is the longest possible match. The function *maxmatch*(*s*) takes a sequence as input, which is split into morphs. It uses the parameters *i* and *j* that specify the beginning and end of the morph, respectively. First, it is assumed that the entire word is a morph, and if there are no coincidences, the position *j* is decreased by one. If the substring under consideration coincides with some morph on the list, the morph is marked. The boundary *i* is moved and placed at the end of the marked morph, and the boundary *j* is again placed at the end of the word. The procedure continues until the boundary *i* reaches the end of the word.

3.4 *Log_likelihood* Model

The model is based on finding the maximum likelihood for morphs. All possible combinations of morpheme boundaries in the word under analysis are considered, then those are selected in which the resulting word segments can occur at a given position and are found in the list of morphs. Then the logarithms of the probabilities of the candidate analyses are calculated. The analysis that has the maximum value is selected.

By processing all morphemic analyses from Tikhonov's dictionary [13], an associative array of morphs, positions and frequencies is created. Then a list of all possible word segmentations is obtained. In total, 2^{x-1} ways of segmenting the word are possible, where x is the length of the word in characters (it is assumed that the boundary cannot occur at the beginning of the word). For each of these segmentations, a bit sequence mask is created that contains information about the presence of morphemic boundaries: if there is no morpheme boundary after the symbol, then the value in the corresponding position of the mask is 0, and if there is a boundary, then the value is 1. From all segmentations, the system selects the ones in which the resulting word segments are all found in the morph list.

Next, the most probable morphemic analysis of the word is selected. For each candidate analysis, the system computes the product of the probabilities of the morphs occurring in the word. The analysis with the highest value of the natural logarithm of this product is chosen as the most probable one. Since the number of possible segmentations is exponential in the length of the word, the *maxmatch* model is used instead of *log_likelihood* for words longer than 18 characters.

3.5 Arithmetic Mean Model (*mean)*

This is a slight modification of the *log_likelihood* model. The *mean* model also considers all possible segmentations of the word under analysis, but computes the arithmetic mean of morph probabilities for each candidate analysis. The one for which this arithmetic mean is greatest is chosen as the best analysis.

3.6 Combined Models

These models are combinations of the above ones: *rules_corrected*, *maxmatch*, *log_likelihood*, and *mean*. First, the *rules_corrected* model extracts postfixes, inflections, prefixes and suffixes. For finding the root and suffixes not found by *rules_corrected*, one of the three other models (*maxmatch*, *log_likelihood*, or *mean*) is used.

3.7 Morphemic Annotation of Text

All models allow for two modes of operation: analysis of individual words or text annotation. When the text annotation mode is chosen, the system simply performs the analysis or each successive token in the text. Function words and interjections are skipped. For the *maxmatch*, *log_likelihood*, *mean*, *rules_corrected+maxmatch*, *rules_corrected+log_likelihood*, and *rules_corrected+mean* models, annotation only includes morphemic boundaries within every word. For the *rules* and *rules_corrected* models, morph types are also part of the annotation. The *rules* and *rules_corrected* models also make it possible to search for a morph by its type.

4 Evaluation

4.1 Evaluation Metrics

We use the metrics precision, recall, and F-measure in the same form as they were applied to morphemic analysis evaluation by K. Ak and O.T. Yildiz [22]. The values of these metrics are calculated based on the following parameters: *hits* is the number of correct boundaries (true positives), *insertions* is the number of unnecessary boundaries (false positives), and *deletions* is the number of overlooked boundaries (false negatives).

Then precision, recall, and F-measure are calculated as follows:

$$Precision = \frac{hits}{hits + insertions}$$

$$Recall = \frac{hits}{hits + deletions}$$

$$F - measure = \frac{2 \times hits}{2 \times hits + insertions + deletions}$$

4.2 Evaluation Setting

For evaluation, a random sample of 500 words from the Tikhonov dictionary was obtained. It is important to note that these words had been removed from the dictionary before any training of our models took place. Thus, the 500 words were completely 'new' to all our models. This test set was used to compare the performance of our models described in Section 3 with one of the available morphemic analysis tools for Russian [15] that we set as the baseline. The latter returns several candidate analyses. For testing, we chose the analysis listed as the most probable.

4.3 Results and Discussion

Table 1. Evaluation Results.

Model	Precision	Recall	F-score
rules	0.905	0.639	0.749
rules_corrected	**0.944**	0.63	0.756
maxmatch	0.73	0.567	0.638
log_likelihood	0.73	0.567	0.638
mean	0.652	0.795	0.716
rules_corrected + maxmatch	0.846	0.85	**0.848**
rules_corrected + log_likelihood	0.847	0.847	0.847
rules_corrected + mean	0.551	**0.915**	0.687
External system [15]	0.834	0.713	0.769

The *rules* model shows very high precision (0.905) since it takes into account word forms and derivational connections, so it is very accurate at marking postfixes, inflections, prefixes and form-building suffixes. The small recall value (0.639) can be explained by the absence of capabilities for analyzing complex words, as well as ignoring some of the word-building suffixes and finding non-existent prefixes. The *rules_corrected* model demonstrates even better precision (0.944) due to more accurate prefix-finding.

The *maxmatch*, *log_likelihood* and *mean* models yield lower results, since they do not take into account form-building and derivational connections between words. The low recall (0.567) of the *maxmatch* model is due to the fact that the model tries to match the longest possible morphs, and the relatively high recall (0.795) of the *mean* model can be explained by the high frequency of short morphs, which leads this model to segmenting words into smaller parts. The *maxmatch* and *log_likelihood* models have the same metric values, because the product of probabilities increases with the decrease in the number of factors, which corresponds to a decrease in the number of morphs.

The F-measure values of the *rules_corrected+maxmatch* (0.848) and *rules_corrected+log_likelihood* (0.847) models are quite high due to the fact that these models produce morphemic segmentation for a larger number of derivational suffixes and complex words. These models decisively outperform the existing morphemic analysis system [15] set as the baseline (0.769).

5 Conclusion and Future Work

In an effort to solve the morphemic analysis task for the Russian language, we have developed a few models: rule-based, probabilistic and combined. The *rules* model was created based on the rules of form building and word formation. The improved version of the model, *rules_corrected*, has better precision due to more accurate marking of prefixes. The *maxmatch*, *log_likelihood*, and *mean* models use such characteristics of morphs as length, frequency and position. By combining the *rules_corrected* and *maxmatch* models we managed to achieve the best performance of 0.848 F-measure on a gold standard set of 500 held-out words analyzed for morphemic structure.

Our system, which is made available to the community [23], also features morphemic annotation of arbitrary texts and morpheme search. The best-performing models take into account the form-building patterns, derivational connections between words, the part of the speech of the analyzed word and its other morphological features. The system can analyze previously unseen and complex words, as well as words in non-initial forms.

As is usually the case, there is still room for improvement. We believe that even better quality of morphemic analysis is achievable by paying more attention to word-formative suffixes and improving the model for analyzing complex words. In terms of functionality, it is also possible to implement search for related words in a text.

References

1. Cotterell, R., Schütze, H.: Joint Semantic Synthesis and Morphological Analysis of the Derived Word. In: Transactions of the Association for Computational Linguistics, vol 6, pp 33-48 (2018)
2. Fritzinger, F., Fraser, A.: How to Avoid Burning Ducks: Combining Linguistic Analysis and Corpus Statistics for German Compound Processing. In: Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, pp 224-234 (2010)
3. Sennric, R., Haddow, B., Birch, A.: Neural Machine Translation of Rare Words with Subword Units. In: Proceedings of the 59th ACL, pp 1715-1725 (2016)
4. Plungyan, V.A.: General Morphology: the introduction into subject [Obshchaya morfologiya: Vvedenie v problematiku]: textbook, 2nd edition, 384 pp. Moscow: Editorial URSS (2003)
5. Huckvale, M., Fang, A.: Experiments in Applying Morphological Analysis in Speech Recognition and Their Cognitive Explanation. In: IOA Conference on Speech and Hearing. http://discovery.ucl.ac.uk/74330/
6. Karpov, A.A.: Models and program realization of Russian speech recognition based on morphemic analysis [Modeli i programmnaya realizatsiya raspoznavaniya russkoy rechi na osnove morfemnogo analiza], a PhD thesis. Saint-Petersburg, 129 pp (2007)
7. Tachbelie, M., Abate S., Menzel W.: Morpheme-based Automatic Speech Recognition for A Morphologically Rich Language Amharic. In: Proceedings of the 2nd International Workshop on Spoken Languages Technologies for Under-resourced Languages (SLTU'10), pp 68 - 73 (2010)
8. Tagabileva, M.G., Berezutskaya, Yu.N.: Word-formation annotation of the Russian National Corpus: aims and methods [Slovoobrazovatel'naya razmetka Natsional'nogo Korpusa russkogo yazyka: zadachi i metody]. In: Computational Linguistics and Intellectual Technologies Papers from the Annual International Conference "Dialogue" [Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Po materialam ezhegodnoy Mezhdunarodnoy konferentsii «Dialog»], issue 9 (16), pp 499 - 507. Moscow: Russian State University for the Humanities [Rossiyskiy gosudarstvennyy gumanitarnyy universitet] (2010)
9. Kukushkina, O.V.: The problems of morphemic decomposition and automatization of the process of morphemic segmentation of the Russian word [Problemy morfemnogo chleneniya i avtomatizatsiya protsessa morfemnoy segmentatsii russkogo slova]. In: Russian computational and quantitative linguistics [Russkaya komp'yuternaya i kvantitativnaya lingvistika]. http://philol.msu.ru>~rlc2001...files/komp_linv/doc
10. Decomposition of words [Razbor slov po sostavu]. http://www. morphemeonline
11. Search in the dictionaries. Morphemics [Poisk v slovaryakh. Morfemika]. http://www.udarenieru.ru
12. Online dictionaries [Slovari onlayn]. http://www.slovonline.ru
13. Dictionaries of the Russian language for downloading. The archives of the forum "Speak Russian" [Slovari russkogo yazyka dlya skachivaniya. Arkhivy foruma «Govorim po-russki»]. http://www.speakrus.ru/dict/
14. Russian Grammar [Russkaya grammatika]. Vol 1: Phonetics. Phonology. Accent. Intonation. Word-formation. Morphology [Fonetika. Fonologiya. Udarenie. Intonatsiya. Slovoobrazovanie. Morfologiya] / N.Yu. Shvedova (main editor), 789 pp. Moscow: Science [Nauka] (1980)
15. Morphological, phonetic and morphemic analysis of the word [Morfologicheskiy, foneticheskiy i morfemnyy razbor slova]. https://vnutrislova.net

16. Ronzhin, A.L., Karpov, A.A.: Implementation of morphemic analysis for Russian speech recognition. In: 9th Conference Speech and Computer, 2004. http://www.isca-speech.org/archive_open/specom_04/spc4_291.pdf
17. Edush, M.V., Dikiy, P.V.: The algorithm and the practical realization of the morphemic decomposition [Algoritm i prakticheskaya realizatsiya morfemnogo razbora]. http://taac.org.ua/files/a2011/proceedings/RU-1-Dikiy%20Petr%20Viktorovich-82.pdf
18. Fadeev, S.G., Zheltov, P.V.: Optimization options of word forms morphemic analysis on the basis of statistical knowledge. In: Russian linguistic bulletin 3 (7), pp 15 - 17 (2016)
19. The Russian National Corpus. The search in the corpus: the main corpus [Natsional'nyy korpus russkogo yazyka. Poisk v korpuse: osnovnoy korpus]. http://www.ruscorpora.ru/search-main.html
20. Bernhard, D.: Unsupervised Morphological Segmentation Based on Segment Predictability and Word Segments Alignment. In: M. Kurimo, M. Creutz, and K. Lagus (eds.), Proceedings of the Pascal Challenges Workshop on the Unsupervised Segmentation of Words into Morphemes, pp 19-23 (2006)
21. Sapin, A.S., Bol'shakova, E.I.: Features of the construction of morphoprocessor CrossMorphy for the Russian language [Osobennosti postroeniya morfoprotsessora russkogo yazyka CrossMorphy]. In: New information technologies in automatic systems: materials of the 20th workshop [Novye informatsionnye tekhnologii v avtomatizirovannykh sistemakh: materialy dvadtsatogo nauchno-prakticheskogo seminara]. Moscow: Keldysh Institute of Applied Mathematics [Institut prikladnoy matematiki imeni M. V. Keldysha], pp 73 - 81 (2017)
22. Ak, K., Yildiz O.T.: Unsupervised morphological analysis using tries. In: Computer and Information Sciences II. – Springer, London, pp 69-75 (2011)
23. Morphemic analysis system for Russian. https://github.com/LudmilaMaltina/morphemic-analysis-rus