Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2018"

Moscow, May 30—June 2, 2018

MORPHOLOGICAL SEGMENTATION WITH SEQUENCE TO SEQUENCE NEURAL NETWORK

Arefyev N. V. (narefjev@cs.msu.su)[1]
Lomonosov Moscow State University, Moscow, Russia;
Samsung Moscow Research Center, Moscow, Russia

Gratsianova T. Y. (tgratsianova@cs.msu.su),
Popov K. P. (kpopov94@ya.ru)[1]
Lomonosov Moscow State University, Moscow, Russia

Morphological segmentation is an important task of natural language processing as it can significantly improve the processing of unfamiliar and rare words in different tasks that involve text data. In this paper we present datasets in English and Russian for learning and evaluating morphological segmentation algorithms, demonstrate the method based on the sequence to sequence neural model and show that the proposed approach shows better results in comparison with other existing methods of morpheme segmentation. We start from an English dataset, which is already available and only minor preprocessing has been made, and then we experiment with the Russian language, where we could not obtain prepared data. So, some more serious preprocessing issues are included. Moreover, we demonstrate how morphological segmentation can improve another natural language processing task—evaluation of words semantic similarity. To achieve this goal, first we try to reproduce the best results of the participants of Russian words semantic similarity competition (RUSSE), which was conducted in Dialogue 2015 conference. Then we show how with the help of smart morpheme segmentation these results can be advanced.

Keywords: morphological segmentation, sequence transduction, sequence to sequence, semantic similarity

[1] These two authors contributed equally to this work

МОРФЕМНАЯ СЕГМЕНТАЦИЯ С ПОМОЩЬЮ SEQUENCE TO SEQUENCE НЕЙРОННОЙ СЕТИ

Арефьев Н. В. (narefjev@cs.msu.su)[2]
Московский Государственный Университет
им. Ломоносова, Москва, Россия;
Samsung Moscow Research Center, Москва, Россия

Грацианова Т. Ю. (tgratsianova@cs.msu.su),
Попов К. П. (kpopov94@ya.ru)[2]
Московский Государственный Университет
им. Ломоносова, Москва, Россия

1. Introduction

The importance of automatic morphological segmentation lies in the fact that it improves the processing of rare and unknown words, which are common in natural texts. Usually the algorithm designed for solving this task takes words of some language as input, and returns as output the same words, but segmented into morphemes.

Nowadays there are many systems for natural language processing, which are trained on huge amounts of text data. Taking a word for a minimal language unit is a common approach in such algorithms. However, it is known that, for example, in Russian there is an order of magnitude more words than the morphemes. Here by the term *morpheme* we mean the minimal meaningful part of the word. That is why the model that was trained on morphemes instead of words will be much smaller. This fact will allow using such model on devices with limited memory.

Another advantage of using morphemes in natural language processing tasks is the possibility of handling unknown and rare words. For example, if we want to evaluate semantic similarity of two unknown words, we can split them into morphemes and somehow estimate similarity between these morphemes. The fact that morphemes are minimal meaningful parts of words guarantees that such evaluation will be justified.

In this paper we describe the possibility of morphological segmentation with sequence to sequence (seq2seq) model and evaluate results of this approach by different methods. Our main contributions are the following.

1. We adapt sequence to sequence neural network for morphological segmentations and show its superiority over existing models on Russian and English datasets.

[2] Эти два автора внесли одинаковый вклад в эту работу

2. We develop a new dataset for the Russian language for training and evaluation the methods of morphological segmentation. It is described in the section where we present other used datasets.
3. We show that our approach improves the semantic similarity estimation of unknown words.

We compare our method with the existing universal algorithm and with the algorithm developed specially for Russian language, xMorphy[3].

We open sourced our code to facilitate further research in morphological segmentation and it's applications for the Russian language[4].

2. Morphological Segmentation as Sequence Transduction

Sequence to sequence (seq2seq) is a general-purpose neural network architecture for sequence transduction, which is used for tasks such as machine translation, text summarization, conversational modeling and more as described in [Denny Britz et al., 2017]. In this work, we adapt seq2seq, consisting of an encoder and decoder, with an attention mechanism for morphological segmentation. The next three paragraphs briefly describe the encoder, decoder and the attention mechanism.

An encoder reads in «source data», e.g. a sequence of symbols, and produces a vector containing information about this data relevant for the task. The idea is that the representation produced by the encoder can be used by the decoder to generate correct output (solve the task).

A decoder is a generative model that is conditioned on the representation created by the encoder. For example, a Recurrent Neural Network decoder may learn to generate the translation of the encoded sentence into another language.

Instead of encoding the input sequence into a single fixed size representation, the model can, with attention mechanism [Bahdanau et al., 2014], learn how to generate an input representation for each output time step. In other words, the model learns which elements of the input sequence to attend to in order to generate the next element of output sequence, based on the input sequence and what it has produced so far.

For training the model, morphological segmentation task was defined as sequence transduction, that is, the sequence of symbols is being transformed into another sequence of symbols. For this purpose, every word in training datasets was represented as the sequence of its letters, e.g. _б|е|з|о|к|о|н|н|ы|й_ (_w|i|n|d|o|w|l|e|s|s_). Additionally, the special symbol "*" was added into target training dataset. This symbol indicated the boundaries between word's segments, e.g. _б|е|з|*|о|к|о|н|*|н|*|ы|й_ (_w|i|n|d|o|w|*|l|e|s|s_).

Hyperparameters, which we used in training process, are described in the **Table 1** (the same hyperparameters were used in every experiment). The values were taken from authors' recommendations for the amount of data which is close to our [Denny B. et al., 2017].

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

[4] https://github.com/kpopov94/morpheme_seq2seq

Table 1. Seq2seq training hyperparameters

Name	Value	Description
Attention hyperparameters		
num_units	256	Hidden state dimension.
Encoder hyperparams		
num_units	256	Size of the LSTM cell in the encoder
dropout_input_keep_prob	0.8	Apply dropout to the (non-recurrent) inputs of each GRU layer using this keep probability.
num_layers	1	Number of GRU layers.
Decoder hyperparameters		
num_units	256	Size of the LSTM cell in the decoder
dropout_input_keep_prob	0.8	Apply dropout to the (non-recurrent) inputs of each GRU layer using this keep probability.
num_layers	2	Number of GRU layers.
Other hyperparameters		
embedding.dim	256	Dimensionality of the embedding layer.

This model is fully described in [Denny Britz et al., 2017] where the schematic picture of its' operation can be found.

3. Related Work

To date, quite a large number of algorithms have been developed for automatic morpheme segmentation. Basically, such tools use approaches based on the principle of maximum likelihood as in [Creutz M. et al., 2004], MAP [Creutz M. et al., 2007], FSA [Goldsmith J. et al., 2004] and CRF[5].

Until 2010, the annual MorphoChallenge competition was held, where different algorithms of morpheme segmentation were compared. In 2010, a program called Morfessor, which is based on the maximum a posteriori estimation principle showed the best results[6]. Later, in 2013, the Morfessor 2.0 was developed and showed much better results than its predecessor improving F-measure from 60% to 80%. The main innovation of this algorithm was the possibility of learning on both labeled and unlabeled data. It was also possible to set hyperparameters for balancing the importance between the labeled and the unlabeled data to several thousands, as the algorithm authors do for the best results. For example, if the annotated corpus was relatively small, it was necessary to increase the beta coefficient, which was responsible for the weight of labeled data during training.

In this paper, we compare our method to Morfessor 2.0, because it shows better quality than Morfessor 1.0 which was the winner in MorphoChallenge 2010, as was mentioned above.

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

[6] http://morpho.aalto.fi/events/morphochallenge2010/comp1-results.shtml

4. Evaluation

There are two main approaches to morphological segmentation evaluation. In direct evaluation the results of an algorithm are compared to gold standard. Indirect evaluation shows the benefits of predicted morphological segmentations on some other task.

4.1. Direct Evaluation

For the direct evaluation we chose Boundary Precision and Recall (BPR)[7] metric, which was used in MorphoChallenge competition since 2005. This choice was mainly motivated by the fact that the authors of Morfessor 2.0 also used it.

As in many other evaluation methods, precision, recall and F-measure for single word are calculated by these formulas:

$$precision = \frac{number\ of\ correct\ boundaries\ found}{total\ number\ of\ boundaries\ found}$$

$$recall = \frac{number\ of\ correct\ boundaries\ found}{total\ number\ of\ correct\ boundaries}$$

$$F - measure = 2 * \frac{precision * recall}{precision + recall}$$

Here the term *boundary* means the borderline between word's segments, which the algorithm succeeds or fails to discover. For example, there is one boundary in the word *пере*езд* (*re*location*). If we assume, that algorithm segmented this word as *пер*е*езд* with two boundaries, we can calculate precision as 0.5, recall as 1.0 and F-measure for this word will be 0.66.

For calculating total precision, recall and F-measure simple average value are taken through them.

4.2. Indirect Evaluation

As a task for indirect evaluation of morphological segmentation, we decided to use the task of Russian words Semantic Similarity Estimation (RUSSE) introduced as shared task in Dialogue 2015 conference [Panchenko A. et al., 2015]. The results of comparison of three algorithms for solving this task are presented in [Arefyev N. V. et al., 2015]. The best results were shown by the word2vec model and that's why we used it in our experiments.

Word2vec is designed for training on large text data for representing the words as relatively low dimensional dense vectors. This approach allows estimating words similarity using dot product between corresponding vectors. The main disadvantage of this model, perhaps, is the absence of predictions about words that were not included in the model and for rare words the estimation will be totally incorrect.

[7] http://morpho.aalto.fi/events/morphochallenge/software/bpr.py

To solve this problem, the method for resolving the absence of out-of-vocabulary words (denoted as OOV) was proposed. If the next word cannot be found in the model during the evaluation process, this method tries to find some part of this word in the model. That is, the search is done by sequential separation of prefixes from word, letter by letter.

For the demonstration and evaluation of morphological segmentation we have improved this approach. Instead of separating leading symbols, the word is segmented into morphemes with the help of seq2seq model and then the resulted segments and their sequential concatenations are searched in the model. For example, let us assume that the word *жертвовательница* (*sacrificer*) is not in the word2vec model and is segmented as *жертв*ова*тель*ниц*а*. The next processing of this word can be shown in the Table 2 (simplified for demonstration):

Table 2. Word processing in morpheme segmentation resolution approach

window	segments	found in model
1	*жертв* *ова* *тель* *ниц* *а*	none
2	*жертвова* *ователь* *тельниц* *ница*	none
3	*жертвователь* *овательниц* *тельница*	*жертвователь*
4	*жертвовательниц* *овательница*	none

The following estimation of semantic similarity will be done for the word *жертвователь* (*sacrifier*). To be short, hereinafter this method will be referenced as MSR (morpheme segmentation resolution).

As the conclusion for this section, it is worth to mention that we are not solving the word embedding task in this work, but we use this task for evaluation of morphological segmentation quality.

4.3. Datasets

For the experiments in this paper we used several different datasets. Some of the data was preprocessed for lowercasing, replacing letters “ё” to “е”, deleting extra symbols and so on. For seq2seq model training data was represented in the form described in Section 2. In the following description of the data we use the term “word type” to denote unique words in the data and the term “token” to denote a particular occurrence of a word:

1. All English text data came from MorphoChallenge 2010 competition[8]:
    - 878,036 unsegmented word types for training
    - 1,000 segmented word types for training
    - 686 segmented word types for testing
2. Lib.rus.ec book collection—424,362 unsegmented words for training.
3. [Tikhonov A. N. 2008]—98,186 segmented word types, that were used in experiments both for training and evaluation[9].
4. Russian Wikipedia—238,052,379 tokens, that were used in RUSSE competition in 2015. We used this corpus for training word2vec model for repeating results from [Arefyev et al., 2015] in Evaluation on RUSSE task section and for experiments with our approach in this task.
5. Datasets from RUSSE competition. They are HJ, RT, AE and AE2 datasets and fully described in [Panchenko A. et al., 2015].

For the training and testing purposes, words from Tikhonov were randomly divided into train and test sets in 3:1 proportion.

5. Results

5.1. BPR on English datasets

For initial evaluation of the proposed method, we decided to compare it with Morfessor on English dataset, which was used in [Sami Virpioja et al., 2013]. We used the same trainsets and the same hyperparameters for training Morfessor as in [Sami Virpioja et al., 2013].

Seq2seq was trained with hyperparameters described in **Section 2**.

We used 1000 segmented words for training every algorithm and we also used 878,036 unsegmented words for Morfessor training.

The difference between original results for Morfessor, presented in [Sami Virpioja et al., 2013], and our reproduced results on the same training dataset can be explained by the different test sets. The authors of Morfessor 2.0 have not mention where it is possible to get the test set they used in experiments in their paper, so we just used the remaining words (i.e. the words have not been used for training) from MorphoChallenge 2010 dataset.

Results are shown in the **Table 3**:

[8] http://morpho.aalto.fi/events/morphochallenge2010/datasets.shtml#download

[9] Russian dataset for seq2seq training and evaluation available at https://drive.google.com/file/d/1DcAVZ4Nv5Xbeua8vnW4SUylnxGqvR5yn/view

Table 3. Results for English dataset

Method	Test data	Precision	Recall	F-measure
Morfessor [Sami Virpioja et al., 2013]	1,000	0.8591	0.8550	0.8571
Morfessor (reproduced)	686	0.8676	0.8530	0.8603
Seq2seq	686	0.9019	0.8716	0.8865

As we can see, seq2seq model shows both better precision and better recall compared to Morfessor 2.0, even though it could not exploit large amount of unsegmented data.

We also compared our approach to MORSE, one of the recently developed algorithms for morphological segmentation. It is fully described in [Tarek Sakakini, 2017]. The authors of MORSE published only one trained model for English, and no source code for training were provided.

Test data were truncated to 539 words, because for 147 words MORSE did not provide predictions, which are equal to their origin words, e.g. *accompanied* turned to *accompany|ed*. The BPR evaluation method that we used does not accept such situations. Results are in the **Table 4**.

Table 4. Morfessor, MORSE and seq2seq comparison

Method	Precision	Recall	F-measure
Morfessor	0.8663	0.8719	0.8691
MORSE	0.9125	0.6785	0.7783
Seq2seq	0.8914	0.8922	0.8918

Seq2seq again showed the best results, and even Morfessor demonstrated its superiority against MORSE. The difference between MORSE and seq2seq is more than 0.1.

5.2. BPR on Russian datasets

Models for Russian words were trained on Tikhonov dictionary [Tikhonov A. N., 2008] as described in **Section 4.3**. For Morfessor training we used unsegmented lib.rus.ec corpus, which is also described in **Section 4.3**.

Morfessor was trained with hyperparameters from previous section and seq2seq was trained with parameters, which were described in **Section 2**.

We also compared seq2seq model and Morfessor 2.0 to xMorphy[10] tool, which is the only tool we found for morphological segmentation of words in Russian. Unfortunately, the authors do not supply code for training so we could not train it on our data.

We used 73,639 segmented words for seq2seq and Morfessor training and also 424,362 unsegmented words for Morfessor. For test we took the remaining part of Tikhonov dictionary which was 24,547 words.

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

Table 5. Results for Russian dataset

Method	Precision	Recall	F-measure
xMorphy	0.7797	0.6589	0.7143
Morfessor	0.9127	0.8904	0.9014
Seq2seq	0.9407	0.9383	0.9395

We can see from results that seq2seq gives 4% and 22% better F-measure than Morfessor and xMorphy respectively.

Considering the fact that BPR applies only the segmentation boundaries by evaluation algorithm, there are counts for different types of errors in the Table 6 for the demonstration purposes. While counting errors, the segmentation was considered correct only if it fully coincided with the gold standard, otherwise it was considered incorrect.

Table 6. Comparison of count of errors done by two algorithms

	Morfessor correct	Morfessor incorrect
seq2seq correct	13,806 words	5,396 words
seq2seq incorrect	2,695 words	2,650 words

As we can see, seq2seq shows better results than Morfessor. And if we take a closer look on separate words, we can notice that errors by seq2seq were made with relatively difficult words. There are some random examples in Table 7:

Table 7. Random examples of complicated segmentations

Word	Seq2seq result	Gold standard segmentation
венчиковый	вен*чик*ов*ый	вен*ч*ик*ов*ый
забытье	за*бы*ть*е	забы*ть*е
статичный	статич*н*ый	стат*ич*н*ый
расточать	рас*точ*а*ть	расточ*а*ть
скрыться	с*кры*ть*ся	скры*ть*ся
шилоклювка	шил*о*клюв*к*а	ши*л*о*клюв*к*а
подержаться	подерж*а*ть*ся	по*держ*а*ть*ся

5.3. Evaluation on RUSSE task

In the evaluation on RUSSE task we used the same seq2seq model as we did in previous section, and benchmarks described in [Arefyev et al., 2015].

Four semantic similarity approach were done (first two are the same as in the [Arefyev et al., 2015]):

1. Without out-of-vocabulary optimization.
2. With out-of-vocabulary optimization (OOV).
3. With MSR optimization.
4. With MSR and OOV (where MSR failed) optimization.

Results are in the Table 8:

Table 8. RUSSE evaluation results

Method	HJ	RT	AE	AE2
No optimization [Arefyev et al., 2015]	0.53200	0.73100	0.88100	0.91400
No optimization (reproduced)	0.52964	0.73563	0.88310	0.91253
OOV (reproduced)	0.55314	0.81217	0.91381	0.91909
MSR	0.56845	0.82738	0.91448	0.91941
MSR + OOV	0.56845	0.82849	0.91507	0.92039

A little difference in first two rows of the table can be explained by differences in the text preprocessing before word2vec training.

As we can see from the results, the method with morphological segmentation shows better results for every evaluation dataset. The Table 9 contains the words for which MSR was able to improve word similarity evaluation (without optimization the value would be zero). Unknown words are marked by "?" signs.

Table 9. Examples of MSR improvements

First word	Second word	First word + MSR	Second word + MSR	Words similarity measure
осмысление	?осмысливание?	осмысление	мысл	0.62
авиасообщение	?авиауслуга?	авиасообщение	авиа	0.48
аудио	?аудиопродукция?	аудио	аудио	1.00
?ахвахец?	?ахвахска?	ахвах	ахвах	1.00
сатана	?люциферов?	сатана	люцифер	0.61

6. Conclusion and Further Work

After all of experiments, the proposed model demonstrates a sustainable superiority over Morfessor 2.0. We also proved that the morphemic segmentation can improve the results of word semantic similarity estimation.

A relatively small improvement on the RUSSE problem can be developed by improving the algorithm for finding concatenations of morphemes in the model due to some limitations, for example, you can try to limit the size of segments for which search will be made and others not to be considered, and also focus only on options with root morphemes.

As we said earlier, the representation of natural language as units in the form of vectors for solving natural language processing problems is quite common. If we take words as units, several problems immediately arise:

1. The model produces a large number of vectors, which directly affects the size of the model.
2. For rare words, vectors are unrepresentative, that is, the vector representation of a rare word does not carry in itself useful data and does not reflect reality.

3. Since natural language is a constantly expanding system, it is inevitable that many words of natural language will not enter the model anyway, no matter how large is the amount of data for training.

On the other hand, letters could be taken as language units. Then it would take only, for example, 33 vectors for the Russian language, since there are only 33 letters in it. This would significantly reduce the volume of the model, but the practical use of this approach would be extremely small, since letters do not carry semantic information. It is known that morphemes are the minimal meaningful parts of words. The language has much less morphemes than words. This fact allow this representation to solve the problem of a large volume of the model. The problem of rare words will be greatly simplified due to the fact that even rare words consist of morphemes, which, at least some of them, will be recognized by the model. The disadvantage of this approach is an increase in ambiguity.

References

1. *Arefyev N. V., Panchenko A. I., Lukanin A. V., Lesota O. O., Romanov P. V.* (2015), Evaluating Three Corpus-based Semantic Similarity Systems for Russian, available at: http://www.dialog-21.ru/media/1119/arefyevnvetal.pdf
2. *Bahdanau D., Cho K., Bengio Y.* (2014), Neural Machine Translation by Jointly Learning to Align and Translate, available at http://www.cl.uni-heidelberg.de/courses/ws14/deepl/BahdanauETAL14.pdf
3. *Creutz M., Lagus K.* (2004), Induction of a simple morphology for highly-inflecting languages. In Proceedings of th7th Meeting of the ACL Special Interest Group in Computational Phonology (SIGPHON). Barcelona, Spain. 43–51.
4. *Creutz M., Lagus, K.* (2007), Unsupervised models for morpheme segmentation and morphology learning. ACM Trans. Speech Lang. Process. 4, 1, Article 3.
5. *Denny Britz, Anna Goldie, Minh-Thang Luong, Quoc Le* (2017), Massive Exploration of Neural Machine Translation Architectures, available at: arxiv.org/pdf/1703.03906.pdf
6. *Goldsmith J., Hu Y.* (2004), From signatures to finite state automata. Midwest Computational Linguistics Colloquium, available at: https://newtraell.cs.uchicago.edu/files/tr_authentic/TR-2005-05.pdf
7. *Panchenko A., Loukachevitch N. V., Ustalov D., Paperno D., Meyer C. M., Konstantinova N.* (2015), RUSSE: The First Workshop on Russian Semantic Similarity, In proceeding of Dialogue 2015 conference, available at: http://www.dialog-21.ru/media/1123/panchenkoaetal.pdf
8. *Sami Virpioja, Peter Smit, Stig-ArneGrönroos, Mikko Kurimo* (2013), Morfessor2.0: Python Implementation and Extensions for Morfessor Baseline, Aalto University publication series, available at: https://aaltodoc.aalto.fi/bitstream/handle/123456789/11836/isbn9789526055015.pdf
9. *Tarek Sakakini, Suma Bhat, Pramod Viswanath* (2017), MORSE: Semantically Driven MORpheme SEgment-er, available at: https://arxiv.org/pdf/1702.02212.pdf
10. *Tikhonov A. N.* (2008), Morpheme-spelling dictionary of the Russian language [Morfemno-orfograficheskij slovar' russkogo yazyka], ACT, Moscow, Russia.