# Evaluation of Morphological Embeddings for English and Russian Languages

**Vitaly Romanov**
Innopolis University, Innopolis, Russia
v.romanov@innopolis.ru

**Albina Khusainova**
Innopolis University, Innopolis, Russia
a.khusainova@innopolis.ru

## Abstract

This paper evaluates morphology-based embeddings for English and Russian languages. Despite the interest and introduction of several morphology-based word embedding models in the past and acclaimed performance improvements on word similarity and language modeling tasks, in our experiments, we did not observe any stable preference over two of our baseline models - SkipGram and FastText. The performance exhibited by morphological embeddings is the average of the two baselines mentioned above.

## 1 Introduction

One of the most significant shifts in the area of natural language processing is to the practical use of distributed word representations. Collobert et al. (2011) showed that a neural model could achieve close to state-of-the-art results in Part of Speech (POS) tagging and chunking by relying almost only on word embeddings learned with a language model. In modern language processing architectures, high quality pre-trained representations of words are one of the major factors of the resulting model performance.

Although word embeddings became ubiquitous, there is no single benchmark on evaluating their quality (Bakarov, 2018), and popular intrinsic evaluation techniques are subject to criticism (Gladkova and Drozd, 2016). Researchers very often rely on intrinsic evaluation, such as semantic similarity or analogy tasks. While intrinsic evaluations are simple to understand and conduct, they do not necessarily imply the quality of embeddings for all possible tasks (Gladkova et al., 2016).

In this paper, we turn to the evaluation of morphological embeddings for English and Russian languages. Over the last decade, many approaches tried to include subword information into word representations. Such approaches involve additional techniques that perform segmentation of a word into morphemes (Arefyev N.V., 2018; Virpioja et al., 2013). The presumption is that we can potentially increase the quality of distributional representations if we incorporate these segmentations into the language model (LM).

Several approaches that include morphology into word embeddings were proposed, but the evaluation often does not compare proposed embedding methodologies with the most popular embedding vectors - Word2Vec, FastText, Glove. In this paper, we aim at answering the question of whether morphology-based embeddings can be useful, especially for languages with rich morphology (such as Russian). Our contribution is the following:

1. We evaluate simple SkipGram-based (SG-based) morphological embedding models with new intrinsic evaluation BATS dataset (Gladkova et al., 2016)

2. We compare relative gain of using morphological embeddings against Word2Vec and FastText for English and Russian languages

3. We test morphological embeddings on several downstream tasks other than language modeling, i.e., mapping embedding spaces, POS tagging, and chunking

The rest of the paper is organized as follows. Section 2 contains an overview of existing approaches for morphological embeddings and methods of their evaluation. Section 3 explains embedding models that we have tested. Section 4 explains our evaluation approaches. Section 5 describes results.



*Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 77–81
Minneapolis, USA, June 6, 2019. 

## 2 Related work

The idea to include subword information into word representation is not new. The question is how does one obtain morphological segmentation of words. Very often, researchers rely on the unsupervised morphology mining tool Morfessor (Virpioja et al., 2013).

Many approaches use simple composition, e.g., sum, of morpheme vectors to define a word embedding. Botha and Blunsom (2014) were one of the first to try this approach. They showed a considerable drop in perplexity of log-bilinear language model and also tested their model on word similarity and downstream translation task. The translation task was tested against an n-gram language model. Similarly, Qiu et al. (2014) tweak CBOW model so that besides central word it can predict target morphemes in this word. Final embeddings of morphemes are summed together into the word embedding. They test vectors on analogical reasoning and word similarity, showing that incorporating morphemes improves semantic similarity. El-kishky et al. (2018) develop their own morpheme segmentation algorithm and test the resulting embeddings on the LM task with SGNS objective. Their method achieved lower perplexity than FastText and SG.

A slightly different approach was taken by Cotterell and Schütze (2015) who optimized a log-bilinear LM model with a multitask objective, where the second objective is to guess the next morphological tag. They test resulting vector similarity against string distance (morphologically close words have similar substrings) and find that their vectors surpass Word2Vec by a large margin.

Bhatia et al. (2016) construct a hierarchical graphical model that incorporates word morphology to predict the next word and then optimize the variational bound. They compare their model with Word2Vec and the one described by Botha and Blunsom (2014). They found that their method improves results on word similarity but is inferior to approach by Botha and Blunsom (2014) in POS tagging.

Another group of methods tries to incorporate arbitrary morphological information into embedding model. Avraham and Goldberg (2017) observe that it is impossible to achieve both high semantic and syntactic similarity on the Hebrew language. Instead of morphemes, they use other linguistic tags for the word, i.e., lemma, the word itself, and morphological tag. Chaudhary et al. (2018) took the next level of a similar approach. Besides including morphological tags, they include morphemes and character n-grams, and study the possibility of embedding transfer from Turkish to Uighur and from Hindi to Bengali. They test the result on NER and monolingual machine translation.

Another approach that deserves being mentioned here is FastText by Bojanowski et al. (2017). They do not use morphemes explicitly, but instead rely on subword character n-grams, that store morphological information implicitly. This method achieves high scores on both semantic and syntactic similarities, and by far is the most popular word embedding model that also captures word morphology.

There are also approaches that investigate the impact of more complex models like RNN and LSTM. Luong et al. (2013) created a hierarchical language model that uses RNN to combine morphemes of a word to obtain a word representation. Their model performed well on word similarity task. Similarly, Cao and Rei (2016) create Char2Vec BiLSTM for embedding words and train a language model with SG objective. Their model excels at the syntactic similarity.

## 3 Embedding techniques

In this work, we test three embedding models on English and Russian languages: SkipGram, FastText, and MorphGram. The latter one is similar to FastText with the only difference that instead of character n-grams we model word morphemes. This approach was often used in previous research.

All three models are trained using the negative sampling objective

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-m\le j\le m, j\neq 0}\log\sigma(s(w_j,w_t))+ \sum_{i=1}^{k}E_{w\sim P_n(w_t)}\left[\log\sigma(s(w,w_t))\right] \quad (1)$$

In the case of SG, the similarity function $s$ is the inner product of corresponding vectors. FastText and MorphGram are using subword units. We use the same approach to incorporate subword information into the word vector for both models:

$$s(w_j, w_t) = \sum_{s \in \mathcal{S}_{w_t}} v_s^T v_{w_j}$$

where $S_{w_t}$ is the set of word segmentations into n-grams or morphemes. We use Gensim[1] as the implementation for all models (Řehůřek and Sojka, 2010). For MorphGram, we utilize FastText model and substitute the function that computes character n-grams for the function that performs morphological segmentation.

## 4 Experiments and Evaluation

To understand the effect of using morphemes for training word embeddings, we performed intrinsic and extrinsic evaluations of SG, FastText, and MorphGram model for two languages - English and Russian. Russian language, in contrast to English, is characterized by rich morphology, which makes this pair of languages a good choice for exploring the difference in the effect of morphology-based models.

### 4.1 Data and Training Details

We used the first 5GB of unpacked English and Russian Wikipedia dumps[2] as training data.

For training both SG and FastText we used Gensim library, for MorphGram - we adapted Gensim's implementation of FastText by breaking words into morphemes instead of n-grams, all other implementation details left unchanged. Training parameters remain the same as in the original FastText paper, except the learning rate was set to 0.05 at the beginning of the training, and vocabulary size was constrained to 100000 words. Morphemes for English words were generated with polyglot[3], and for Russian - with seq2seq segmentation tool[4].

When reporting our results in tables, we will refer for FastText as FT and MorphGram as Morph.

### 4.2 Similarity

One of the intrinsic evaluations often used for word embeddings is a similarity test - given word pairs with human judgments of similarity degree for words in each pair, human judgments are compared with model scores—the more is the correlation, the better model "understands" semantic similarity of words. We used SimLex-999 (Hill et al., 2015) dataset—the original one for English and its translated by Leviant and Reichart (2015) version for Russian, for evaluating trained embeddings. Out-of-vocabulary words were excluded from tests for all models. The results are presented in Table 1.

| | SG | FT | Morph |
|---|---|---|---|
| en | **0.37** | 0.35 | 0.36 |
| ru | **0.24** | 0.19 | 0.19 |

Table 1: Correlation between human judgments and model scores for similarity datasets, Spearman's $\rho$.

| | | SG | FT | Morph |
|---|---|---|---|---|
| en | Google Semantic | **65.34** | 48.75 | 57.52 |
| | Google Syntactic | 55.88 | **75.10** | 61.16 |
| | BATS | 29.67 | **33.33** | 32.71 |
| ru | Translated Semantic | **39.11** | 25.59 | 34.69 |
| | Translated Syntactic | 32.71 | **59.29** | 43.68 |
| | Synthetic | 24.52 | **36.78** | 27.06 |

Table 2: Accuracy of models on different analogies tasks.

We see that SG beats the other two models on similarity task for both languages, and MorphGram performs almost the same as Fasttext.

### 4.3 Analogies

Another type of intrinsic evaluations is analogies test, where the model is expected to answer questions of the form A is to B as C is to D, D should be predicted. For English, we used Google analogies dataset introduced by Mikolov et al. (Mikolov et al., 2013a) and BATS collection (Gladkova et al., 2016). For Russian, we used a partial translation[5] of Mikolov's dataset, and a synthetic dataset by Abdou et al. (2018).

Again, we excluded all out-of-vocabulary words from tests. We report accuracy for different models in Table 2.

Interestingly, MorphGram is between SG and FastText in semantic categories for both languages, and between FastText and SG for syntactic categories for English.

[1] https://radimrehurek.com/gensim

[2] https://dumps.wikimedia.org/

[3] https://polyglot.readthedocs.io/en/latest/index.html

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

[5] https://rusvectores.org/static/testsets/

| | SG | FT | Morph |
|---|---|---|---|
| ru-en *1-nn* | **56.27** | 55.58 | 53.51 |
| ru-en *10-nn* | **78.96** | 78.82 | 77.03 |

Table 3: Accuracy of supervised mapping from Russian to English using different models, searching among first and ten nearest neighbors.

### 4.4 Mapping Embedding Spaces

Here we introduce a new type of evaluation—it focuses on a cross-lingual task of mapping two embedding spaces for different languages. The core idea is to transform embedding spaces such that after this transformation the vectors of words in one language appear close to the vectors of their translations in another language. We were interested to see if using morphemes has any benefits to perform this kind of mapping.

We map embeddings using a train seed dictionary (dictionary with word meanings) and state of the art supervised mapping method by Artetxe et al. (2018), and calculate the accuracy of the mapping on the test dictionary. In short, the essence of this method is to find optimal orthogonal transforms for both embedding spaces to map them to a shared space based on a seed dictionary, plus some additional steps such as embeddings normalization. For each model—SG, FastText, and MorphGram, we mapped Russian and English embeddings trained using this model. We used the original implementation[6] for mapping (supervised option), and ground-truth train/test dictionaries provided by Facebook for their MUSE[7] library. We report 1-nn and 10-nn accuracy: whether the correct translation was found as a first nearest neighbor or among 10 nearest neighbors of a word in the mapped space. See the results in Table 3.

We observe no positive impact of using MorphGram for mapping word embedding spaces.

### 4.5 POS Tagging and Chunking

Other tasks where incorporation of morphology can be crucial are the tasks of POS Tagging and chunking. We use a simple CNN-based architecture introduced in (Collobert et al., 2011), with one projection layer, one convolutional layer, and the final logit layer. The only input features we use are the embeddings from corresponding models. The English language embeddings are tested with Conll2000 dataset which contains 8935 training sentences and 44 unique POS tags. The dataset for the Russian language contains 49136 sentences and 458 unique POS tags. Due to time constraint, we train models only for a fixed number of epochs: 50 for English and 20 for Russian (iterations reduced due to a larger training set). The results for POS and chunking are given in Tables 4 and 5 correspondingly. It is interesting to note that SG embeddings perform better for English on POS task, but for Russian, embeddings that encode more syntactic information always perform better.

| | SG | FT | Morph |
|---|---|---|---|
| en | **0.9824** | 0.9754 | 0.9722 |
| ru | 0.8817 | **0.8899** | 0.8871 |

Table 4: Accuracy on POS task

| | SG | FT | Morph |
|---|---|---|---|
| en | 0.8966 | **0.9034** | 0.8985 |
| ru | 0.8442 | **0.8548** | 0.8534 |

Table 5: Accuracy on Chunk task

## 5 Results

In this paper, we compared three word embedding approaches for English and Russian languages. The main inquiry was about the relevance of providing morphological information to word embeddings. Experiments showed that morphology-based embeddings exhibit qualities intermediate between semantic driven embedding approaches as SkipGram and character-driven one as FastText. Morphological embeddings studied here showed average performance on both semantic and syntactic tests. We also studied the application of morphological embeddings on two downstream tasks: POS tagging and chunking. For English language, SG provided the best results for POS, whereas FastText gave the best result on chunking task. For Russian, FastText showed better performance on both tasks. Morphological embeddings, again, showed average results. We recognize that the difference in the results on downstream task can be considered marginal. We also did not observe improvements from morphological embeddings on word similarity dataset compared to other models.

[6] https://github.com/artetxem/vecmap

[7] https://github.com/facebookresearch/MUSE

## References


Mostafa Abdou, Artur Kulmizev, and Vinit Ravishankar. 2018. Mgad: Multilingual generation of analogy datasets. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Popov K.P. Arefyev N.V., Gratsianova T.Y. 2018. 24rd International Conference on Computational Linguistics and Intellectual Technologies.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Oded Avraham and Yoav Goldberg. 2017. The interplay of semantics and morphology in word embeddings. *arXiv preprint arXiv:1704.01938*.

Amir Bakarov. 2018. A survey of word embeddings evaluation methods. *arXiv preprint arXiv:1801.09536*.

Parminder Bhatia, Robert Guthrie, and Jacob Eisenstein. 2016. Morphological priors for probabilistic neural word embeddings. *arXiv preprint arXiv:1608.01056*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Jan Botha and Phil Blunsom. 2014. Compositional morphology for word representations and language modelling. In *International Conference on Machine Learning*, pages 1899–1907.

Kris Cao and Marek Rei. 2016. A joint model for word embedding and word morphology. *arXiv preprint arXiv:1606.02601*.

Aditi Chaudhary, Chunting Zhou, Lori Levin, Graham Neubig, David R. Mortensen, and Jaime G. Carbonell. 2018. Adapting Word Embeddings to New Languages with Morphological and Phonological Subword Representations.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

Ryan Cotterell and Hinrich Schütze. 2015. Morphological word-embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1287–1292.

Ahmed El-kishky, Frank Xu, Aston Zhang, Stephen Macke, and Jiawei Han. 2018. Entropy-Based Subword Mining with an Application to Word Embeddings. pages 12–21.

Anna Gladkova and Aleksandr Drozd. 2016. Intrinsic evaluations of word embeddings: What can we do better? In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 36–42.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: What works and what doesn't. In *Proceedings of the NAACL-HLT SRW*, pages 47–54, San Diego, California, June 12-17, 2016. ACL.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Ira Leviant and Roi Reichart. 2015. Judgment language matters: Multilingual vector space models for judgment language aware lexical semantics. *CoRR, abs/1508.00106*.

Thang Luong, Richard Socher, and Christopher Manning. 2013. Better word representations with recursive neural networks for morphology. *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 104—-113.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Siyu Qiu, Qing Cui, Jiang Bian, Bin Gao, and Tie-Yan Liu. 2014. Co-learning of word representations and morpheme representations. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 141–150.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. `http://is.muni.cz/publication/884893/en`.

Sami Virpioja, Peter Smit, Stig-Arne Grönroos, Mikko Kurimo, et al. 2013. Morfessor 2.0: Python implementation and extensions for morfessor baseline.