Evaluation of Sentence Embedding Models for Natural Language Understanding Problems in Russian

Dmitry Popov, Alexander Pugachev, Polina Svyatokum, Elizaveta Svitanko, and Ekaterina Artemova

National Research University Higher School of Economics
{dgpopov, avpugachev, posvyatokum, eisvitanko}@edu.hse.ru,
echernyak@hse.ru

Abstract. We investigate the performance of sentence embeddings models on several tasks for the Russian language. In our comparison, we include such tasks as multiple choice question answering, next sentence prediction, and paraphrase identification. We employ FastText embeddings as a baseline and compare it to ELMo and BERT embeddings. We conduct two series of experiments, using both unsupervised (i.e., based on similarity measure only) and supervised approaches for the tasks. Finally, we present datasets for multiple choice question answering and next sentence prediction in Russian.

Keywords: multiple choice question answering · next sentence prediction · paraphrase identification · sentence embedding

1 Introduction

With word embeddings have been in the focus of researchers for several decades, sentence embeddings recently started to gain more and more attention. A word (sentence) embedding is a projection in a vector space of relatively small dimensionality, that can capture word (sentence) meaning by making the embeddings of two words (sentences) that are similar to get closer in this vector space. With no doubts, the usage of the properly trained word embeddings boosted the quality of the majority of natural language processing (NLP), information extraction (IE), neural machine translation (NMT) tasks. However, it seems that word embedding models are facing their limits when it comes to polysemy and ambiguity. A natural solution to these problems lies in sentence embedding models too, as they allow to capture the context–dependent meaning of any part of the sentence.

For the last two years, the amount of projects and papers on sentence embeddings has increased dramatically. Several research groups show that a complex pre–trained language model may serve both as an input to another architecture and as a standalone sentence embedding model. The most famous models of this type, ELMo, and BERT, named after characters of Sesame Street show, can be treated as "black boxes", that read a sentence in and output a vector representation of the sentence. The efficiency of these models for the English language is

well studied already not only in several natural language understanding (NLU) tasks but also for language modeling and machine translation. However little has been done to explore the quality of sentence embeddings models for other languages, including Russian, probably due to the absence of NLU datasets. The contribution of this paper is two–fold. **First**, we create two novel NLU datasets for the Russian language: a) multiple choice question answering dataset, which consists of open domain questions on various topics; b) next sentence prediction dataset, that can be treated as a kind of multiple choice question answering. Given a sentence, one needs to choose between four possible next sentences. Correct answers are present in both datasets by design, which makes supervised training possible. **Second**, we evaluate the quality of several sentence embedding models for three NLU tasks: multiple choice question answering, next sentence prediction and paraphrase identification.

The results confirm, that in the tasks under consideration, sentence–level representations perform better than the word–level ones as in many other tasks.

The remainder is structured as follows. Section 2 presents an overview of related works; Section 3 introduces the datasets, Section 4 describes the methods we used to tackle the tasks. The results of the experiments are presented in Section 5. Section 6 concludes.

2 Related work

A word embedding is a dense vector representation of a word that allows modeling some sort of semantic (or functional) word similarity. Two words are considered similar if a similarity measure, such as cosine function, for example, between corresponding vectors is high enough. As this definition is rather vague, there are two main approaches to evaluating the quality of a word embedding model. Intrinsic evaluation is based on conducting on standard word pairs and analogies datasets, such as Word–353[1] or Simlex–999[2]. External evaluation requires an external machine learning tasks, such as sentiment classification or news clustering, which can be evaluated by a quality measure, such as accuracy or Rank index. All factors of machine learning algorithms are held equals so that these quality measures are affected by the word embeddings model only.

The methodology of creation and evaluation of sentence embeddings is less developed, when compared to the zoo of word embedding models. The evaluation of sentence embeddings models is usually conducted of natural language understanding tasks, such as semantic textual similarity, natural language inference, question answering, etc. Further, we overview the basic and more advanced models of sentence embeddings.

2.1 Unsupervised sentence embeddings

The simplest way of obtaining the sentence embedding is by taking the average of the word embeddings in the sentence. The averaging can treat words equally, rely

[1] `http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/`
[2] `https://fh295.github.io/simlex.html`

on $tf - idf$ weights [1], take the power mean of concatenated word embeddings [17] and employ other weighting and averaging techniques.

Another approach of unsupervised training of sentence embeddings is Doc2Vec [12], which succeeds after Word2Vec and FastText. Word2Vec [13] by Mikolov et al. is a predictive embedding model and has two main neural network architectures: continuous Bag–of–Words (CBoW) and continuous skip–gram [14]. Given a central word and its context (i.e., k words to the left and k words to the right), CBoW tries to predict the context words based on the central one, while skip–gram on tries to predict the context words based on the central one.

Joulin et al. [9] suggest an approach called FastText, which is built on Word2Vec by learning embeddings for each subword (i.e., character n–gram, where n can vary between some bounds and is a hyperparameter to the model). To achieve the desired word embeddings, the subword embeddings are averaged into one vector at each training step. While this adds a lot of additional computation to training, nevertheless it enables word embeddings to encode sub–word information, which appears to be crucial for morphologically rich languages, the Russian language being one of them.

The goal of the aforementioned Doc2Vec approach, introduced by Mikolov et al. [12], is to create an embedding of a document, regardless of its length.

All the studies as mentioned earlier work properly on the English language and some of them release pre–trained Russian embeddings, too. Russian–specific RusVectores [11] pre–trained model, which was trained on Russian National Corpora, Russian Wikipedia and other Web corpora possesses several pre–trained word embeddings models and allows to conduct a meaningful comparison between the models and their hyperparameters.

The Skip–Thought model follows the skip–gram approach: given a central sentence, the context (i.e., the previous and the next) sentences are predicted. The architecture of Skip–Thought consists of a single encoder, which encodes the central sentence and two decoders, that decode the context sentences. All three parts are based on recurrent neural networks. Thus their training is rather difficult and time–consuming.

2.2 Supervised sentence embeddings

In recent years, several sources say the unsupervised efforts to obtain embeddings for larger chunks of text, such as sentences, are not as successful as the supervised methods. Conneau et al. [5] introduced the universal sentence representation method, which works better than, for instance, Skip–Thought [10] on a wide range of transfer tasks. It's model architecture with BiLSTM network and max–pooling is the best current universal sentence encoding method. The paper on Universal Sentence Encoder [3] discovers that transfer learning using sentence embeddings, which tends to outperform the word level transfer. As an advantage, it needs a small amount of data to be trained in a supervised fashion.

2.3 Language models

One of the recently introduced and efficient methods are embeddings from Language Models (ELMo) [16] that models both complex characteristics of word use, and how it is different across various linguistic contexts and can also be applied to the whole sentence instead of the words. Bidirectional Encoder Representations from Transformers (BERT) [6] has recently presented state–of–the–art results in a wide variety of NLP tasks including Natural Language Inference, Question Answering, and others. The application of an attention model called Transformer allows a language model to have a better understanding of the language context. In comparison to the single–direction language models, this one uses another technique namely Masked LM (MLM) for a bidirectional training.

2.4 Evaluation of sentence embedding models

While different embedding methods are previously discussed, the most suitable evaluation metric is also a challenge. Quality metrics are largely overviewed in RepEval[3] proceedings. In [2] both widely–used and experimental methods are described. SentEval is used to measure the quality of sentence representations for the tasks of natural language inference or sentence similarity [4]. The General Language Understanding Evaluation (GLUE) benchmark[4] is one of the popular benchmarks for evaluation of natural understanding systems. The top solutions, according to the GLUE leadership, exploit some sort of sentence embedding frameworks. GLUE allows testing any model in nine sentence or sentence–pair tasks, such as natural language inference (NLI), semantic textual similarity (paraphrase identification, STS) or question answering (QA).

3 Datasets

3.1 Multiple Choice Question Answering (MCQA)

The portal geetest.ru provides tests on many subjects such as history, biology, economics, math, etc., with most of the tasks being simple wh–questions. These tests were downloaded to create a multiple choice question answering dataset.

Every test is a set of questions in a specific area of a certain subject. For the final dataset we handpicked tests from the following subjects: Medicine, Biology, History, Geography, Economics, Pedagogy, Informatics, Social Studies.

The selection was based on two criteria. Firstly, questions should be answerable without knowing the topic of the test. For example, some questions in test could not be answered correctly without presenting a context of a specific legal system. Secondly, questions should test factual knowledge and not skills. For example, almost any math test will require to perform computations, and such type of task is not suitable for this dataset.

[3] https://aclweb.org/anthology/events/repeval-2017/
[4] https://gluebenchmark.com

What is more, we have collected questions on history and geography from ege.sdamgia.ru. The selection of questions was similar to described above. As a result, the total number of questions is around 11k with three subjects being larger than other. These subjects are: medicine, (4k of questions), history (3k of questions), biology (2k of questions). The resulting dataset is somewhat similar to Trivia QA dataset, however the domains are different [8].

3.2 Multiple choice next sentence prediction (NSP)

We have collected a new dataset with 54k multiple choice questions where the objective is to predict the correct continuation for a given context sentence from four possible answer choices. The dataset was produced using the corpora of news articles of "Lenta.ru"[5]. To sample correct and incorrect answer choices we chose a trigram and a context sentence which ends on this trigram. The correct answer choice was the continuation of the context sentence, and the incorrect choices were the sentences following the trigram in other sentences in the corpora. Labels of correct answers were sampled uniformly. So, the random or constant predictions results in an accuracy score of 0.25.

3.3 Paraphrase identification (PI)

For a paraphrase detection task, we used Russian language paraphrase dataset collected from news titles [6]. The dataset contains 7k pairs of titles which are the same, close and different by meaning. Constant prediction on this dataset gives us an accuracy score of 0.64. In a sense, Microsoft Research Paraphrase Corpora[7] is similar to this dataset. Both of them were collected from news titles.

3.4 Dataset statistics

Frequency distribution of top 25 most frequent tokens, the number of unique tokens and the total number of tokens in the datasets can be found in Figure 1. Sentence length distribution, average and median sentence length can be found in Figure 2.

4 Methods

There are two types of problems we were considered for the comparison of sentence embeddings:

- **Multiple choice questions.** Datasets for this type are MCQA and NSP ones. The objective of the problem is to predict the correct answer for a given question/context and four answer choices.

[5] `https://github.com/yutkin/Lenta.Ru-News-Dataset`
[6] `http://paraphraser.ru`

Question
Какие из указанных симптомов характерны для фарингита?
Answer choices
1. резкая боль в горле
2. першение и дискомфорт в горле
3. затруднение проглатывания слюны
4. субфебрильная температура

Table 1. Examples from MCQA dataset. The correct answer is bolded.

Context
Мартин Скорсезе намеревается приступить к съемкам экранизации романа «Молчание» японского писателя Сюсаку Эндо в 2014 году,
Answer choices
1. был оснащен одиннадцатиметровым стеклянным полом шириной два метра, сообщается в полученном «Домом» пресс-релизе компании «Сен-Гобен».
2. когда он провозил мак для одной из продуктовых баз
3. сообщает Deadline. Финансированием проекта займутся компании Emmett/Furla Films и Corsan Films
4. а производить трубы там начали уже спустя два года. В числе поставщиков «Газпрома» ЗТЗ появился в 2017 году

Table 2. An example from NSP dataset. The correct answer is bolded.

text 1	text 2	label
Мэрилин Мэнсон передумал выступать в России.	Мэрилин Мэнсон отменил тур по России.	1
Бывший чемпион мира по боксу умер в 48 лет.	Как судей судили за их решения.	0

Table 3. Examples from the PI dataset.

– **Paraphrase identification.** The goal for such a problem is to predict if two given sentences are a paraphrase or not.

To compare different methods of obtaining sentence embeddings, we have explored supervised and unsupervised scenarios of using such embeddings.

4.1 Unsupervised approach

In unsupervised methods, we are interested in a similarity between sentence embeddings in terms of cosine similarity.

Multiple choice questions. For such type of problem for a given set $\{q, a_1, a_2, a_3, a_4\}$, where q is an embedding of either a question or a context sentence and a_i is an embedding of $i–th$ answer choice, the predicted answer choice is the choice which embedding is the most similar to q.

Paraphrase identification. Let for a given pair of sentences t_1 and t_2 are sentence embeddings of this pair respectively. First of all, we split the dataset into training and test sets, after that searching for a threshold t on a training

Fig. 1. Frequency distribution of top 25 tokens for MCQA, NSP, PI datasets.

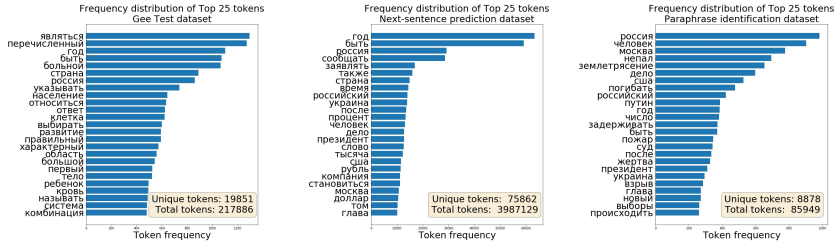
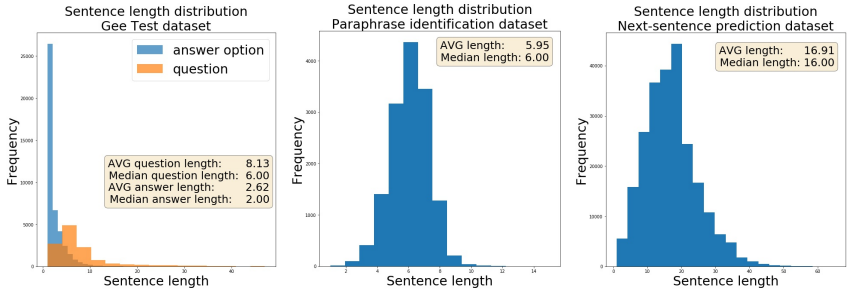


Fig. 2. Sentence length distribution for MCQA, NSP, PI datasets.



set such that pairs with $sim(t_1, t_2) > t$ will be labeled as a paraphrase. Finally, we will evaluate results on a test set.

4.2 Supervised approach

Text vector representations are often inputs to machine learning models. In this approach, we are aiming to figure out which methods of obtaining vector representations are better as inputs into linear models such as the logistic regression and which methods are better as inputs to a gradient boosting models such as the CatBoost [15].

Multiple choice questions. Since we cannot just run a multiclass classification as the correct answers numbers are not related to questions we will make a binary classification model which predicts a probability for a given question–answer (context–answer) pair to be correct, i.e. the answer is the correct answer choice for the question/context. Then for a given question/context and four answers, the predicted answer is the answer such that the model gives the highest probability.

Paraphrase identification. For this problem, we just build a binary classifier on a concatenation of sentence embeddings.

5 Experiments and results

5.1 FastText

Table 4. FastText embeddings results.

Method	MCQA		NSP		PI	
	Accuracy	F1	Accuracy	F1	Accuracy	F1
Unsupervised	0.305	0.304	0.337	0.337	0.719	0.806
Logistic regression	0.287	0.287	0.248	0.248	0.704	0.612
CatBoost	0.318	0.317	0.496	0.496	0.762	0.715

As it can be seen from Table 4, FastText model accomplishes differently depending on tasks and methods. The best quality on each task was reached by CatBoost method according to Accuracy metrics. According to the F1 score, CatBoost also outperforms other methods in all of the tasks except Paraphrase identification where the best score with a large margin was achieved by the unsupervised method. There can also be seen a big quality difference between CatBoost and other methods in next sentence prediction task. Logistic regression shows the worst results in each task. The overall performance of methods based on FastText model is far from perfect. The results achieved using FastText can be considered as a baseline for our investigation.

5.2 ELMo

We have used three ELMo models [7] pre–trained on Russian Wikipedia, "Lenta.Ru" news articles and Russian tweets corpora, respectively. So, one of the main results obtained from our experiments is how different domains of pre–trained models affects final results.

The results achieved by three methods based on three different pre–trained ELMo models are presented in Table 5. The performance of these methods is quite sensitive to the source of training data for ELMo model. For example, regarding the MCQA task, models trained on the News and Twitter corpora perform better than the model trained on Wikipedia, especially when logistic regression is used. In the unsupervised setting, the quality of next sentence prediction task highly depends on the source of training data ELMo model, too. However, in most cases there is no significant difference in results between three ELMo models.

One can notice that logistic regression in both cases (FastText model and all three ELMo models) shows the worst results in the majority of the tasks. Regarding the next sentence prediction task, the performance of logistic regression has not gone far away from random choice. However, it shows better results

[7] `http://docs.deeppavlov.ai/en/master/intro/pretrained_vectors.html`

Table 5. ELMo unsupervised experiments results.

Method	Domain	MCQA		NSP		PI	
		Accuracy	F1	Accuracy	F1	Accuracy	F1
Unsupervised	Wikipedia	0.300	0.300	0.645	0.645	0.807	0.867
	News	0.293	0.293	0.691	0.691	0.807	0.866
	Twitter	0.291	0.291	0.559	0.559	0.803	0.863
Logistic regression	Wikipedia	0.301	0.300	0.249	0.249	0.684	0.652
	News	0.318	0.318	0.249	0.248	0.702	0.668
	Twitter	0.317	0.316	0.251	0.250	0.705	0.674
CatBoost	Wikipedia	0.314	0.314	0.647	0.647	0.773	0.729
	News	0.310	0.310	0.669	0.669	0.797	0.758
	Twitter	0.314	0.314	0.631	0.631	0.779	0.741

than other models in the MCQA task. The unsupervised method achieved the best results for the paraphrase identification task. We can claim that the use of ELMo model contributed to better results in next sentence and paraphrase identification tasks, as there was observed significant improvement in accuracy and F1 scores. Speaking of MCQA task, the results are comparable with the previously obtained.

5.3 BERT

The results of different methods based on BERT embeddings are shown in Tables 6 – 11. All the results were obtained using BERT–Base Multilingual Cased model[8]. There were considered different BERT model layers and combinations of layers. For each method and task, there is presented the best result achieved with layer indication. From Table 6 we can see that using BERT embeddings we can significantly improve results in MCQA task. The CatBoost method based on BERT model noticeably outperforms FastText and ELMo models within this task. According to Table 7, we can claim that BERT model could not get better results compared to ELMo. Even the performance of logistic regression remained the same. However, as we can notice from Table 8, the performance of BERT model regarding paraphrase identification task is comparable to the ELMo results. In both cases, the best score was achieved by the unsupervised method and the worst by logistic regression. We suppose that in many cases BERT could not outperform ELMo model because it was not trained for these tasks and the best way for BERT is to fine–tune it by ourselves.

6 Conclusion

We tested three sentence embedding models: (a) FastText averaged over words in the sentence, b) three pre–trained on various sources ELMo models, c) BERT model in three NLU tasks for the Russian language. These tasks are a) multiple

[8] `https://github.com/google-research/bert/blob/master/multilingual.md`

Table 6. BERT embeddings results on MCQA dataset.

Method	Best score			Average score		Worst score	
	Accuracy	F1	Layer	Accuracy	F1	Accuracy	F1
Unsupervised	0.303	0.302	Concatenation of layers from 1 to 6	0.292	0.292	0.274	0.274
Logistic regression	0.336	0.335	Layer number 1	0.324	0.323	0.310	0.310
CatBoost	0.346	0.346	Max pooling of layers from 4 to 6	0.326	0.326	0.312	0.311

Table 7. BERT embeddings results on NSP dataset.

Method	Best score			Average score		Worst score	
	Accuracy	F1	Layer	Accuracy	F1	Accuracy	F1
Unsupervised	0.508	0.508	Layer number 12	0.457	0.457	0.429	0.429
Logistic regression	0.255	0.255	Max pooling of layers from 7 to 9	0.249	0.248	0.244	0.244
CatBoost	0.514	0.514	Average pooling of layers from 7 to 12	0.479	0.479	0.414	0.414

Table 8. BERT embeddings results on PI dataset.

Method	Best score			Average score		Worst score	
	Accuracy	F1	Layer	Accuracy	F1	Accuracy	F1
Unsupervised	0.801	0.857	Average pooling of layers from 1 to 3	0.787	0.851	0.775	0.843
Logistic regression	0.715	0.676	Average pooling of layers from 7 to 12	0.693	0.651	0.670	0.615
CatBoost	0.778	0.732	Layer number 3	0.763	0.713	0.749	0.694

Table 9. BERT embeddings results on MCQA dataset (1st and 12th layer).

Method	**1st layer**		**12th layer**		**Average pooling**	
	Accuracy	F1	Accuracy	F1	Accuracy	F1
Unsupervised	0.298	0.298	0.296	0.296	0.296	0.296
Logistic regression	0.336	0.335	0.321	0.321	0.326	0.326
CatBoost	0.318	0.318	0.318	0.318	0.329	0.330

Table 10. BERT embeddings results on NSP dataset (1st and 12th layer).

Method	**1st layer**		**12th layer**		**Average pooling**	
	Accuracy	F1	Accuracy	F1	Accuracy	F1
Unsupervised	0.429	0.429	0.508	0.508	0.484	0.484
Logistic regression	0.244	0.244	0.248	0.248	0.247	0.246
CatBoost	0.414	0.414	0.497	0.497	0.514	0.514

Table 11. BERT embeddings results on PI dataset (1st and 12th layer).

Method	**1st layer**		**12th layer**		**Average pooling**	
	Accuracy	F1	Accuracy	F1	Accuracy	F1
Unsupervised	0.795	0.853	0.781	0.850	0.789	0.853
Logistic regression	0.670	0.615	0.704	0.664	0.703	0.662
CatBoost	0.762	0.708	0.758	0.706	0.764	0.718

Table 12. Final results.

Model	**MCQA**		**NSP**		**PI**	
	Accuracy	F1	Accuracy	F1	Accuracy	F1
FastText	0.318	0.317	0.496	0.496	0.762	0.806
ELMo	0.318	0.318	0.691	0.691	0.807	0.867
BERT	0.346	0.346	0.514	0.514	0.801	0.857

choice question answering, b) next sentence prediction, c) paraphrase identification. For the first two tasks, we presented our own new datasets. These datasets are designed as multiple choice questions: given a question / a sentence one need to choose a correct option from four possible answers / continuations.

Our experiments show that the MCQA dataset is much more complicated than the other two datasets. The quality of the results for this task is not as high as for two others. All models perform somewhat similar in next sentence prediction and paraphrase identification tasks. The paraphrase identification dataset is highly unbalanced, and the positive examples are in the minority, which may affect the quality if the results.

Overall, we can claim that we started to evaluate the popular sentence embeddings frameworks in GLUE–like fashion for the Russian language. So far we can state that (1) the word–level embeddings are outperformed by the sentence–level embeddings, (2) the pre–trained models available online with no doubts can attempt some of the NLU tasks with little or almost no fine tuning. The directions of the future work may include probing of embedding models for Russian rich morphology and free word order. The code of all experiments is available on GitHub[9].

Acknowledgements

This project was supported by the framework of the HSE University Basic Research Program and Russian Academic Excellence Project "5–100".

References

1. Arroyo-Fernández, I., Méndez-Cruz, C.F., Sierra, G., Torres-Moreno, J.M., Sidorov, G.: Unsupervised sentence representations as word information series: Revisiting tf–idf. Computer Speech & Language **56**, 107–129 (2019)
2. Bakarov, A.: A survey of word embeddings evaluation methods (2018)
3. Cer, D., Yang, Y., Kong, S.y., Hua, N., Limtiaco, N., John, R.S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., et al.: Universal sentence encoder. arXiv preprint arXiv:1803.11175 (2018)
4. Conneau, A., Kiela, D.: Senteval: An evaluation toolkit for universal sentence representations (2018)
5. Conneau, A., Kiela, D., Schwenk, H., Barrault, L., Bordes, A.: Supervised learning of universal sentence representations from natural language inference data. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 670–680. Association for Computational Linguistics, Copenhagen, Denmark (September 2017), https://www.aclweb.org/anthology/D17-1070
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2018)
7. Dolan, W., Quirk, C., Brockett, C.: Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources. https://www.microsoft.com/en-us/download/details.aspx?id=52398

[9] https://github.com/foksly/aist-sentence-embeddings

8. Joshi, M., Choi, E., Weld, D.S., Zettlemoyer, L.: Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. arXiv preprint arXiv:1705.03551 (2018)
9. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. CoRR **abs/1607.01759** (2016), http://arxiv.org/abs/1607.01759
10. Kiros, R., Zhu, Y., Salakhutdinov, R., Zemel, R.S., Torralba, A., Urtasun, R., Fidler, S.: Skip-thought vectors (2015)
11. Kutuzov, A., Kuzmenko, E.: Webvectors: a toolkit for building web interfaces for vector semantic models. In: International Conference on Analysis of Images, Social Networks and Texts. pp. 155–161. Springer (2016)
12. Le, Q.V., Mikolov, T.: Distributed representations of sentences and documents (2014)
13. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. CoRR **abs/1301.3781** (2013), http://arxiv.org/abs/1301.3781
14. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. CoRR **abs/1310.4546** (2013), http://arxiv.org/abs/1310.4546
15. Ostroumova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A.: Catboost: unbiased boosting with categorical features. arXiv preprint arXiv:1706.09516 (2018)
16. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2018). https://doi.org/10.18653/v1/n18-1202, http://dx.doi.org/10.18653/v1/N18-1202
17. Rücklé, A., Eger, S., Peyrard, M., Gurevych, I.: Concatenated power mean word embeddings as universal cross-lingual sentence representations. arXiv preprint arXiv:1803.01400 (2018)