ATLANTIS PRESS

*Atlantis Studies in Uncertainty Modelling, volume 1*

11th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT 2019)

# An approach to vocabulary expansion for neural network language model by means of hierarchical clustering

**Dudarin Pavel and Yarushkina Nadezhda**
Informational Systems, Faculty of Informational Systems and Technology
Ulyanovsk State Technical University, Severny Venetz st.32
p.dudarin@ulstu.ru, jng@ulstu.ru

## Abstract

Neural network language models become the main tool to solve tasks in NLP field. These models already have shown state-of-the-art results in classification, translation, named entity recognition and so on. Pre-trained models are distributed freely in the internet, and could be reused with help of transfer learning techniques. However, the real life problem's domain could differ from the origin domain which the network was trained. In this paper an approach to vocabulary expansion for neural network language model by means of hierarchical clustering is proposed. This technique allows to adopt prerained language model to a different domain. Firstly, tokens from the language model are hierarchically clustered. Then new words from problem's domain are matched to the tokens accordingly obtained hierarchy. In the experimental part the proposed approach is demonstrated on the slightly modified ULMFiT language model.

**Keywords:** NLP, Language model, Neural Network, RNN, ULMFiT, Clustering, Fuzzy graph clustering, Word-to-vec

## 1 Introduction

Natural language processing (NLP) nowadays is a quickly developing area. NLP finds more and more new fields of application, for example in software analysis [16] and nutrition production [19]. Traditionally NLP tasks deal with long text collected in data sets, there statistic based method could be used. However, with the internet and social media development text tend to become shorter and shorter. A few years ago short text analysis was almost impossible, and only few papers were dedicated to study this problem [5, 3]. Nowadays, processing short texts is becoming a trend [7] in information retrieval [14]. Since the text has rarely external information, it is more challenging than document [20]. In order to solve this task different clustering techniques are used [3, 22]. Each clustering procedure needs a similarity measure [17], and the most used technique to obtain this measure in NLP tasks is word2vec [15].

Although the word embedding approach has shown good efficiency [2], lately an approach of construction neural network language models get a leading position in NLP benchmarks [21], almost every state-of-the-art results are obtaining by means of neural networks. But the process of neural network learning is quite long and computationally expensive.

Besides there are a lot of task in specific domains where there is no opportunity to teach special neural network. In this case the idea of transfer learning [11] looks very promising. Authors of ULMFiT propose using their universal architecture to train language model and then to tune them for specific NLP tasks. But in ULMFiT the tokens list is limited, authors recommend using up to 60 000 tokens. And as long as different word forms are treated as different tokens, ULMFiT's vocabulary is even more limited. On the contrary, modern word embedding models [12] have 250-400 thousand of lemmas. Word embedding technique being combined with thesaurus could demonstrate even higher performance[13]. In case of Russian language with its huge possible word forms language model approach allows to construct general purpose neural networks like casual phases generator only. And does not allow to include specific terms, neologism, swear words, rare used words and so on. In ELMo [4] and BERT [6] words are split into parts and then fed to neural network. But these models take a lot of calculation resources and could be afforded by huge corporations like Google. There are some multilingual pre-trained ELMo [10] and BERT models. But as for now they demonstrate very poor performance for Rus-

sian language. For example 'wish you a Merry Christmas' really common phrase without double meaning could not be continued correctly by available models.

In this paper an approach to customization of pre-trained neural network language model to specific domain is proposed. This technique allows to process word outside the tokens list and thus to get benefits from transfer learning.

The rest of this paper is organized as follows. In section 2 the detailed technique description is presented. Section 3 shows an experimental results. And section 4 concludes the paper.

## 2 Language Model Customization

General idea of proposed approach is to add an extra layer of words pre-processing before the neural network language model (Figure 1).

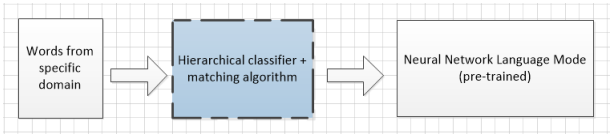


Figure 1: Additional pre-processing layer

This layer consists of two part: hierarchical classifier that groups words from neural network vocabulary and matching algorithm that matches new words to linear combination of words from vocabulary $new\_word = weight_1 * word_1 + ... + weight_N * word_N$.

### 2.1 Tokens Hierarchical Clustering

The first layer of neural network language model is an embedding layer which transforms one-hot encoded vectors into n-dimensional vectors of the embedding vectors space. Each coordinate of one-hot vector references to a word in a vocabulary of language model.

Lets define $W_{lm}$ - a set of words included in tokens list of neural network. The task is to organize words from tokens list into a tree, where leaf nodes contain single word $w_i \in W_{lm}$, and other nodes are clusters that include all the words below in the hierarchy $w_{kj} \in C_k \subset W_{lm}$. $|W_{lm}| = N$.

This task could be completed by performing procedure which is a hierarchical modification[9] of $\epsilon - clustering$ [18, 1]. This procedure needs to be provided with a similarity measure for objects, let denote it as $\mu$. There are a lot of pre-trained word embedding models for each language. This model provide a vector for each word and than the Euclidean or Manhattan distance could be calculated. In this paper the '$ruwikiruscorpora\_upos\_skipgram\_300\_2\_2019$' [1] model was used.

One of the main advantages of graph based approach is its ability to be interpreted by human. The classifier could be easily modified by experts to add information domain specifics [8]. At least all the words that are not from domain vocabulary could be cut of the classifier. On the Figure 2 a part of sample classifier is shown. This sub-tree consist of two main branches dedicated to software and hardware installation process. Each level has a number ($\epsilon$) that indicates the step of hierarchical clustering procedure when these level was obtained and it means that all the branches on this level has mutual similarity less than $\epsilon$.

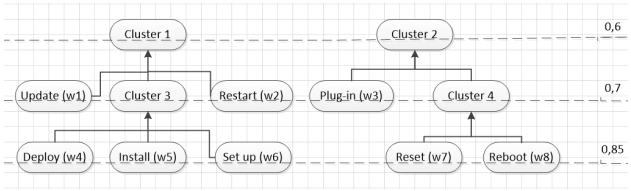


Figure 2: Hierarchical clustering sub-tree sample

Thus a hierarchical classifier with additional layer information could be obtained.

### 2.2 Specific Domain Words Matching

The task of the matching step is to construct vectors for words from specific domain in order they could be processed by pre-trained neural network. These vectors should have N components, where N equals to amount of inputs of neural network $N = |W_{lm}|$ .

For each word $w$ there are two possible cases. The word is already included into language model tokens list $w = w_i \in W_{lm}$ and in this case corresponding vector $v = (0, 0, ..., 1, 0, ..., 0)$, where component with 1 has $i$ index. Another case when word $w \notin W_{lm}$. In this case there are some possible strategies to get a vector form. The first one, and the most evident, is to replace a given word with the most similar one according to similarity measure $\mu$. Means, to choose $i$, $\mu(w, w_i) = max(w, w_j) \forall j \in [1, N]$. This strategy does not require any classifier, but it is not efficient when there are some equidistant words in the tokens list, especially when they are significantly differ in their semantic meaning. In order to have an alternative way of matching, in the experimental part the first strategy also included.

In general case proposed technique is following:

1. If $max(w, w_j) = \mu(w, w_i) > 0,9^2 \ \forall j \in [1, N]$ then

[1] The model was downloaded from open resource *https* : *//rusvectores.org/ru/models*

[2] The value $\epsilon = 0,9$ was obtained by experimental way

$v = (0, 0, ..., 1, 0, ..., 0)$, with 1 on the $i$-th place.

2. Start with $\epsilon = 0,9$ and find all the words $W_{nn} = \{w_j | \mu(w, w_j) >= \epsilon,\ j \in [1, N]\}$. If $|W_{nn}| = 0$ then set $\epsilon = \epsilon - \delta_\epsilon$. In this paper $\delta_\epsilon = 0,05$, according to hierarchical clustering procedure specifics.

3. Get all the clusters $C_{nn} = \{c_j | \exists i\, w_i \in W_{nn}\}$ i.e. all the parent nodes in classifier for leaf nodes in $W_{nn}$.

4. Start with layer $l = 0,9$ and get all nodes from this layer $L_l = \{c_j | c_j \in C_{nn}\, \& \, layer(c_j) = l\}$. If $|L_l| > 2$ [3] then change $l = l + \delta_l$ and move to the previous step. In this paper $\delta_l$ has been chosen as 0,05, according to hierarchical clustering procedure specifics.

5. For each node (cluster) define a weight according the distance to the cluster center. $weight = \mu(w, cluster\ center_i) \,/\, |\sum_{j \in L_l} \mu(w, cluster\ center_j)|$

6. For each child node define weight the same as at the previous step and multiply to parent's weight $weight = parent\ weight * children\ weight$.

7. Stop when all the leaf nodes get weights. All the other weights are set to 0 $weight_i = 0 \forall i \notin W_{nn}$ As a result $v = (weight_1, weight_2, weight_3, ..., weight_N)$

This algorithm is illustrated on Figure 3. Firstly the similarity of word 'mount' to other words is calculated. The most similar words 'install', 'set up' and 'plug-in' were detected. Then layer by layer from the bottom to the top parent nodes are detected, until only 2 nodes left. Next, top-to-bottom process starts. Based on the distance to the cluster centers (0.8 and 0.71), node weights are calculated 0.53 and 0.47 respectively. And finally, weights for children nodes of 'cluster 3' are calculated. Thus a vector for word 'mount' will be (0, 0, 0.53, 0, 0.24, 0.23, 0, 0, ...).

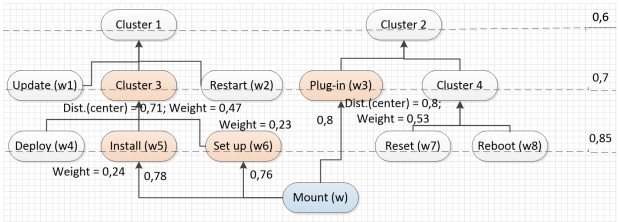


Figure 3: Matching process sample for word 'mount'

Thus each word from the domain is converted into vectors in N-dimensional vector space.

Besides even words that are present in language model token list could be re-matched to another words or set of words. This could be useful in case of word's meaning is changed significantly in the domain. For example: 'mount', 'branch', 'bug' in software development domain.

## 3 Experiment results

For experimental purposes pre-trained neural network language model for Russian language with architecture ULMFiT has been chosen [4]. This model has been trained on news portal (lenta.ru) and has perplexity 36,23.

All the most popular neural network language models take as an input sequence of words, to be more specific - sequence of words indexes in tokens list. This make difficult to use custom input vectors with pre-trained neural network. In this paper hard code solution was used: the 'fastai' library has been modified to change not used input components into hard coded vectors.

To show the technique some common phrases from developers chats is used:

1. 'who can **mount** a new hard drive?'
2. 'this part has a **bug** you need to fix it'
3. 'this abstract class does not satisfy to this **interface**'

The chosen language model does include 'mount', 'bug' in common meaning and does not include word 'interface' in its tokens list. The aim is to be able to proceed this sentences with pre-trained neural network.

The first step is to construct a hierarchical classifier. The input layer of the current network has 60 000 neurons. The resulting hierarchy has about 80 000 nodes, 60 levels. The part of hierarchy is shown on Figure 2.

The second step is to construct vectors for words that are absent in tokens list. Word 'mount' is related to words 'install', 'set up' and 'plug-in'. This case is shown on Figure 3. For the other two words:

1. 'bug': 'failure', 'error', 'lack'
2. 'interface': 'structure', 'rule', 'protocol'

Then the sentences could be processed by neural network language model. The first 3-5 generated words has been taken as an output result:

and could differ from model to model.

[3] this threshold is heuristic and need to be surveyed more thoroughly in future studies

[4] https://github.com/ppleskov/Russian-Language-Model

1. Input: 'who can **mount** a new hard drive?'. Output: 'server has processor core'

2. Input: 'this part has a **bug** you need to fix it'. Output: 'patch will be coming soon '

3. Input: 'this class could not be inherited from this **interface**'. Output: 'protocol is failed'

The results below were generated when one the most similar word has been used instead of vector calculation.

1. Input: 'who can **mount** a new hard drive?'. Output: 'trip will be long and pleasant'

2. Input: 'this part has a **bug** you need to fix it'. Output: 'anti insect service'

3. Input: 'this class could not be inherited from this **interface**'. Output: 'the whole building is a heritage'

Neural network output in first two cases uses the common word meanings and produces wrong context. In the last case the word 'interface' were just ignored and the context produced was based on word 'inherited' only.

## 4 Conclusion

In this paper an attempt to apply transfer learning technique to special domains was made. The proposed approach allows to use not learned words with pretrained neural network language model. It is important in domains with insufficient amount of texts to train custom language model or when the calculation resources are limited. Also this technique could be used to prototype and check ideas (hypothesis) before starting to teach custom language model.

The results shows effectiveness of proposed approach but more thorough experiments need to be done. Further studies will involve comparison of different neural network architectures within proposed approach, searching a way of fine tuning the language model and comparison of effectiveness in different NLP benchmarks. Besides it is important to develop extension to existing neural network frameworks to support not only a custom head but custom tails also.

**Acknowledgement**

The work was supported by the Russian Foundation for Basic Research (Projects No. 17-07-00973, No. 18-47-730019).

## References

[1] R. A., Fuzzy graphs, Fuzzy Sets and Their Applications to Cognitive and Decision Processes. Academic Press, New York. pp. 77–95.

[2] N. Arefyev, P. Ermolaev, A. Panchenko, How much does a word weigh? weighting word embeddings for word sense induction, CoRR abs/1805.09209.

[3] D. P. Avendaño, H. Jiménez-Salazar, P. Rosso, Clustering abstracts of scientific texts using the transition point technique, in: Computational Linguistics and Intelligent Text Processing, 7th International Conference, CICLing 2006, Mexico City, Mexico, February 19-25, 2006, Proceedings, 2006, pp. 536–546.
URL https://doi.org/10.1007/11671299_55

[4] W. Che, Y. Liu, Y. Wang, B. Zheng, T. Liu, Towards better UD parsing: Deep contextualized word embeddings, ensemble, and treebank concatenation, in: Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 55–64.
URL www.aclweb.org/anthology/K18-2005

[5] D. Cohn, R. Caruana, A. Mccallum, Semi-supervised clustering with user feedback, Tech. rep. (2003).

[6] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805.

[7] P. Dudarin, A. Pinkov, N. Yarushkina, Methodology and the algorithm for clustering economic analytics object. Automation of Control Processes, in: Automation of Control Processes, Vol. 47, 1, RGGU, Ulyanovsk, Russia, 2017, pp. 591–604.

[8] P. Dudarin, M. Samokhvalov, N. Yarushkina, An approach to feature space construction from clustering feature tree, in: S. O. Kuznetsov, G. S. Osipov, V. L. Stefanuk (Eds.), Artificial Intelligence, Springer International Publishing, Cham, 2018, pp. 176–189.

[9] P. V. Dudarin, N. G. Yarushkina, An approach to fuzzy hierarchical clustering of short text fragments based on fuzzy graph clustering, in: A. Abraham, S. Kovalev, V. Tarassov, V. Snasel, M. Vasileva, A. Sukhanov (Eds.), Proceedings of the Second International Scientific Conference

"Intelligent Information Technologies for Industry" (IITI'17), Springer International Publishing, Cham, 2018, pp. 295–304.

[10] M. Fares, A. Kutuzov, S. Oepen, E. Velldal, Word vectors, reuse, and replicability: Towards a community repository of large-text resources, in: Proceedings of the 21st Nordic Conference on Computational Linguistics, Association for Computational Linguistics, Gothenburg, Sweden, 2017, pp. 271–276.
URL www.aclweb.org/anthology/W17-0237

[11] J. Howard, S. Ruder, Universal language model fine-tuning for text classification, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2018, pp. 328–339.
URL http://aclweb.org/anthology/P18-1031

[12] A. Kutuzov, E. Kuzmenko, WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models, Springer International Publishing, Cham, 2017, pp. 155–161.

[13] N. Loukachevitch, E. Parkhomenko, Recognition of multiword expressions using word embeddings, in: S. O. Kuznetsov, G. S. Osipov, V. L. Stefanuk (Eds.), Artificial Intelligence, Springer International Publishing, Cham, 2018, pp. 112–124.

[14] C. D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge University Press, New York, NY, USA, 2008.

[15] T. Mikolov, K. Chen, G. S. Corrado, J. Dean, Efficient estimation of word representations in vector space, CoRR abs/1301.3781.

[16] Y. Nadezhda, G. Gleb, D. Pavel, S. Vladimir, An approach to similar software projects searching and architecture analysis based on artificial intelligence methods, in: A. Abraham, S. Kovalev, V. Tarassov, V. Snasel, A. Sukhanov (Eds.), Proceedings of the Third International Scientific Conference "Intelligent Information Technologies for Industry" (IITI'18), Springer International Publishing, Cham, 2019, pp. 341–352.

[17] A. Panchenko, P. Romanov, O. Morozova, H. Naets, A. Philippovich, A. Romanov, C. Fairon, Serelex: Search and visualization of semantically related words, in: P. Serdyukov, P. Braslavski, S. O. Kuznetsov, J. Kamps, S. Rüger, E. Agichtein, I. Segalovich, E. Yilmaz (Eds.), Advances in Information Retrieval, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 837–840.

[18] T. Y. Raymond, S. Bang, Fuzzy relation, fuzzy graphs and their applications to clustering analysis, Fuzzy Sets and their Applications to Cognitive and Decision Pro-cesses. Academic Press. Pages 125-149.

[19] N. Shelekhova, V. Polyakov, E. Serba, S. Tamara, Prospects of application it-technologies in food industry, Nutrition industry 8 (2018) 30–33.

[20] J. Tang, X. Wang, H. Gao, X. Hu, H. Liu, Enriching short text representation in microblog for clustering, Frontiers of Computer Science 6 (1) (2012) 88–101.
URL doi.org/10.1007/s11704-011-1167-7

[21] J. Xu, B. Xu, S. Zheng, G. Tian, J. Zhao, Self-taught convolutional neural networks for short text clustering, Neural networks : the official journal of the International Neural Network Society 88 (2017) 22–31.

[22] Q. Zhao, M. Rezaei, H. Chen, P.: Keyword clustering for automatic categorization, in: In: 2012 21st International Conference on Pattern Recognition (ICPR). IEEE (2012, 2012.