Scientific Visualization, 2022, volume 14, number 4, pages 110 - 120, DOI: 10.26583/sv.14.4.10

Visual Interpretation of Russian Static Vector Space

O.A. Serikov[1], E.S. Klyshinsky[2], V.A. Ganeeva[3]

National Research University "Higher School of Economics",
Myasnitskaya str. 20, 101000, Russia

[1] ORCID: 0000-0002-3746-2642, srkvoa@gmail.com
[2] ORCID: 0000-0002-4020-488X, eklyshinsky@hse.ru
[3] ORCID: 0000-0002-9569-9197, vaganeeva@edu.hse.ru

Abstract

The vector analogy task states that it is possible to find a vector translation which changes a selected semantic feature of a word and connects vector representations of two words. It is well known that the accuracy of such a translation is not always high. In this paper, we introduce a new method of visual representation of static vector embedding space which aims to investigate semantic properties of such a space. The main idea of the method is usage of LSA method for separation of a vector space into semantically homogeneous parts. We also use topic word lists embedded into a static vector space for the sake of visualization of results of such separation. During our experiments, we found out that it is possible to interpret not only small groups of vectors but also the global structure of the whole space. The semantic differences among selected global groups depend on the semantic and pragmatic features of texts used for training the vector model — their genre, style, source, lexis etc. The introduced method can be used for construction of a hierarchical model of a vector space.

Keywords: Static vector space, visual interpretation, LSA.

1. Introduction

As it was demonstrated in [1], the task of word analogy can be solved using a semantic vector space. Formally, the task of word analogy in a vector space can be stated as following. Let $v_{a'}$ and v_a be vector embeddings in a multidimensional semantic space corresponding to words a' and a. In this case, the difference $v_{a'} - v_a$ between two vectors demonstrates a semantic relation between words a' and a; such a relation reflects difference in a set of latent semantic properties of those words. Having a vector v_b for a word b, one can find a word y which reflects the same relation with the word b using simple vector operations: $v_y = v_b + v_{a'} - v_a$ (if there is a word y in the given semantic space).

As it was demonstrated in [2] and [3], it is not always possible to find such an analogy because a word could have several meanings. For example, in case of "king – man + woman = queen", a queen is not always a mighty monarch, who defines inner and foreign policy, but sometimes she is merely a king's wife with a different area of responsibility. On the other hand, authors of [4] demonstrated that accuracy in the word analogy task can be high enough for practical applications.

The word analogy task can be reformulated as following. Let us have a vector m in static vector embedding space; the vector m connects two areas of the same space. Let us have words a', a, b' and b having the same analogy – e.g., male and female names for professional positions or relation between a country and its capital. In this case we can use the following formula:

$$v_{a'} - v_a \approx v_{b'} - v_b \approx m. \quad (1)$$

In case of $v_{a'}$ and $v_{b'}$ are neighbors, i.e. are corresponding to the same small area of embedding space, it follows from (1) that v_a and v_b are neighbors as well. This means that vector m is connecting two small areas in an embedding space sharing the same semantic

features inside group and having the same semantic difference between groups – in our case it is male vs female names for professional positions and countries vs capitals. Note that those areas should be compact and clear, i.e. should contain small amount of semantically related words only.

We found out that in most cases it is true. Fig. 1 demonstrates mutual positions of names of herbs, flowers, bushes and trees embedded using the model Araneum_upos_skipgram (http://rusvectores.org/ru/models/). For the sake of demonstration, we used UMAP which corrupts the original space configuration; that is why some words are moved into a neighbor group. We proved that adding extra dimensions into a figure improves the situation, however it lacks of visual clarity. So, one can see that all plants can be separated by their semantic features: edible are opposed to non-edible, spices to vegetables and other crop plants, lumber trees to fruit-bearing trees.

In this paper, we prove a hypothesis that such a division can be automatically conducted not for local but global groups as well, i.e. not for small semantically heterogeneous groups but for large fragments with the number of elements comparable to the number of words in the considered model.

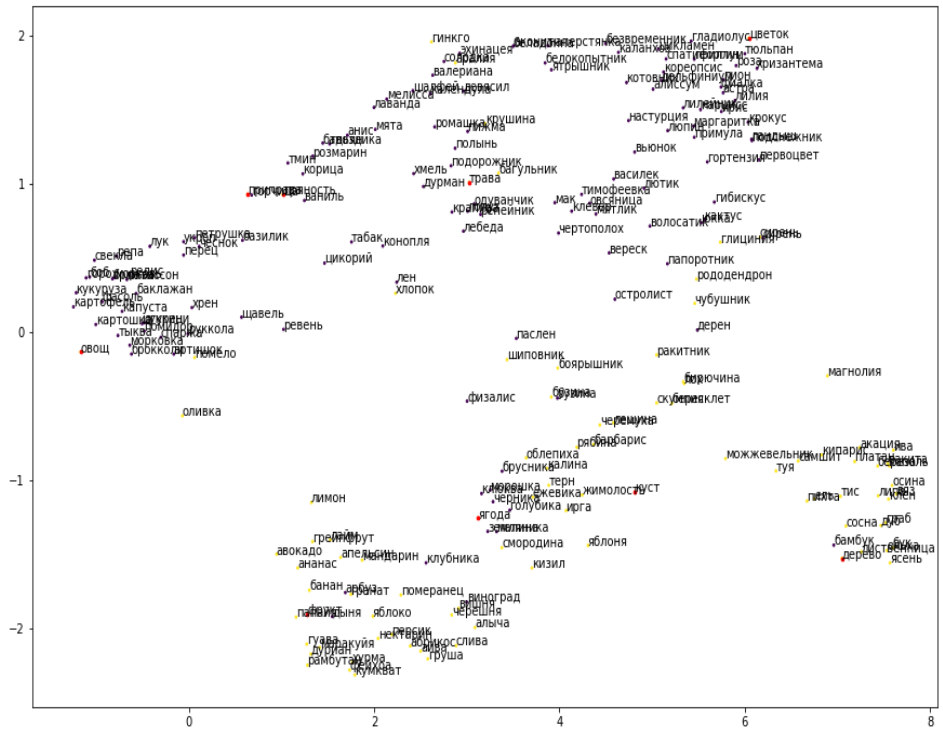

Figure 1 – Separation of words belonging to plants in a vector embedding space

2. Existing Methods for Embedding Space Analysis

One of the modern approaches to the evaluation of semantical properties of language model is probing which considers a model as a black box. Such approach supposes that a researcher passes input vectors to the system and investigates its answers or changes of answers. This approach is used for investigation of both semantic and syntactic properties of a language model. There are three stages of probing [5]: behavioral, diagnostic, and invasive. In case of behavioral probing, a researcher investigates how the model's behavior changes in case of changing of grammatical features of a text; e.g., if a neural network generates a correct text while a word in an input text changes its grammatical number. The diagnostic probing investigates an influence between grammatical feature and model's output. One of the methods here is measuring linear correlation between vector representation of a text and grammatical features in this text. For example, Figure 2 (published in [6]) demonstrates the calculated dependency between the precision for a task and presence in a text one of the

grammatical features. Columns are presenting such tasks in natural language processing as text similarity detection, natural language inference, word classification etc.; rows are presenting such grammatical features as sentence length, depth of a dependency tree, number of subjects or objects in a sentence etc. Finally, the invasive probing adds some noise in a text and investigates "understandability" of this text to a model (see Figure 3). This approach investigates sensitivity of a model to considered grammatic features and if it uses these features at all.

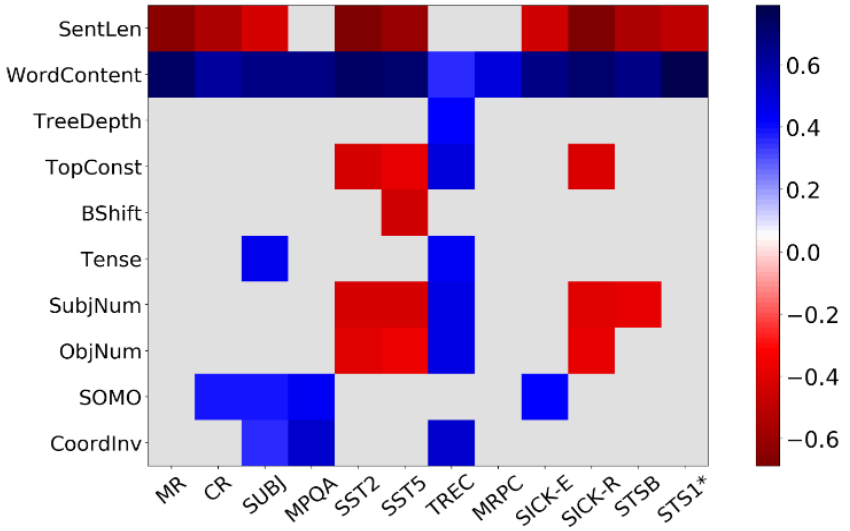

Figure 2 – Correlation matrix between probing features and downstream tasks (cited by [6])

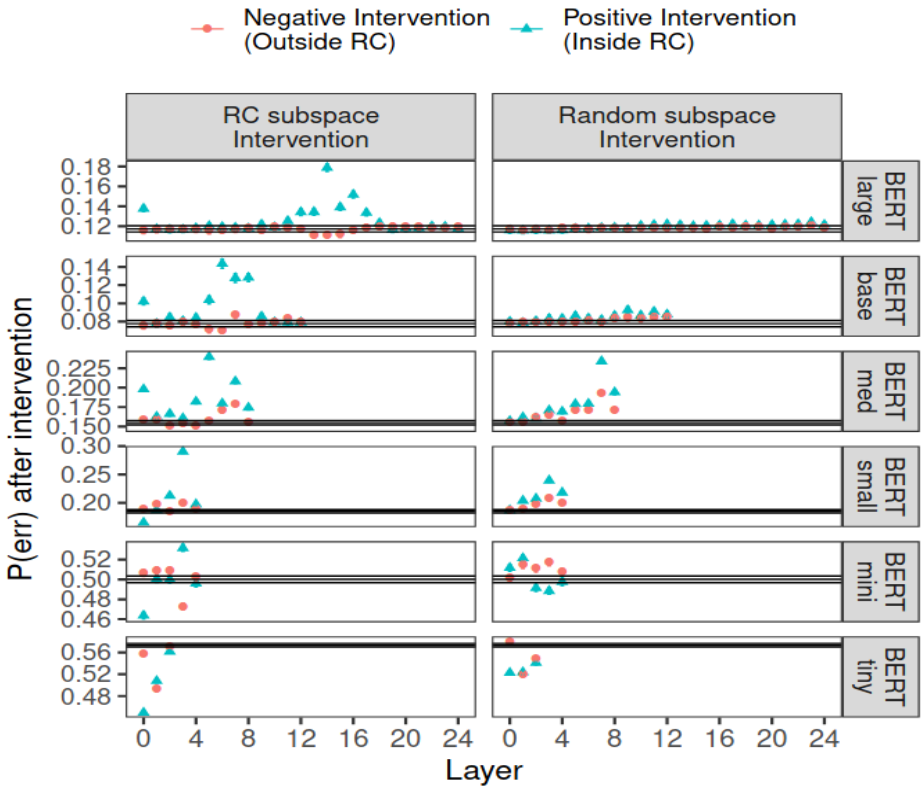

Figure 3 – Change caused by counterfactual representations in agreement error probability across relative clauses with attractors for different BERT variants (cited by [7])

The common way to conduct such investigations is usage of contextualized models. The main difference between contextualized and static models is using a word's context for

construction of an embedding vector. In case of static models, a vector is calculated and assigned to a word independently to its context in a particular case; this loses polysemous and homonymous nature of a word and assigns just one vector for each word. In case of contextualized models, an embedding vector is inferred for every word usage according to the given context. That is why the same word can have different vectors for different contexts. This leads us to the problem of joining different vectors inferred for the similar situations. Note that different word co-occured in a text in neighboring positions could have similar vectors though they have different meanings of belong to different domains.

Investigation of historical meaning shift is another modern actively developed approach. Starting from the first steps in this direction, researchers found out that a word changes its company traveling among domains in course of time. For example, the paper [8] introduces a semantic change detection method using a relative position of a word according to other words. Figure 4 describes an example of word *monumental* which shifted from architecture to informal speech [8]. Currently, there were conducted similar investigations for a variety of languages.

One of the next steps was investigation of mutual position of words depending on their polarity according to a selected semantic feature. The paper [9] demonstrates that usage of names of sports in a text is associated with different affluence – camping and boxing are associated with poor life while golf and tennis with prosperity (see Figure 5).

According to our review, we can state that a Word2Vec embedding vector space has several directions connected to polarity of a semantic feature or a set of such features. The analogy task demonstrates that there are interpretable features or feature sets that are describing semantic changes among group of closely related words. The paper [9] demonstrates that some those directions could be reasonably interpreted by a person.

In this paper we continue our research in area of interpretability of static embedding vector spaces. In our previous papers we proved the interpretability on the local level, using sole words or word groups. In this paper we prove a hypothesis that static vector spaces have some global interpretable directions dividing the vector space at small number huge groups of words which size is comparable to the size of vocabulary of this space.

3. The Method for Analysis of Static Embedding Space

One of the modern methods for composition of closely related groups of words is topic modeling [10]. However, this method has such drawbacks as small number of layers of decomposition and poor interpretability of results. In our research, we use the method of Latent Semantic Analysis (LSA) [11], which is one the most used fundamental method for interpretation of semantic features, in order to investigate its ability to explain a global structure of a vector space. On the one hand, this method allows to create a new latent space which has a better explanation of existed grouping of objects. On the other hand, LSA uses method of Singular Vector Decomposition (SVD) which rotates and scales the original space along axis with the biggest relational deviation. The later provides a better separation of words into closely related groups.

In our research we used static vector embeddings taken from Word2Vec models. As it was mentioned above, these models provide one fixed vector for every word; thus, words' positions can be fixed in the vector space. Usage of contextualized models (e.g. Bert) leads to such problems as careful text selection and clustering and averaging of vectors for different meaning of the same word. Another problem here is interpretation of achieved results of chunking, since Bert models, as well as FastText models, divide a word into fragments, which are hard to interpret without knowing a context of their usage.

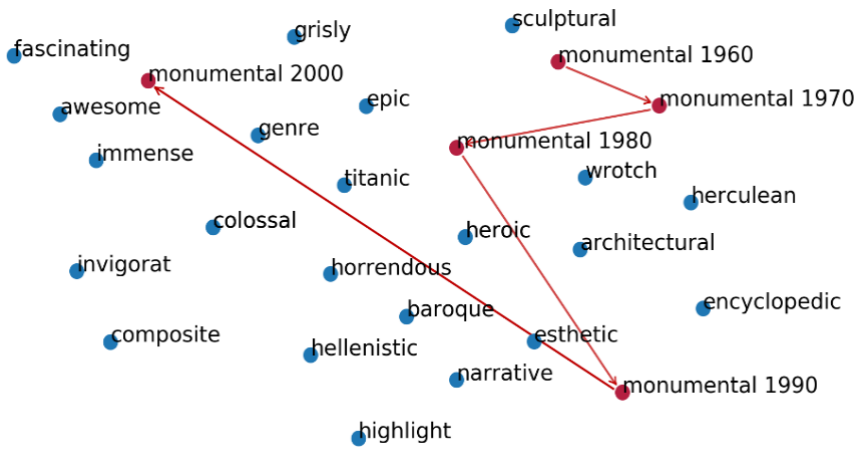

Figure 4 – Alterations in the nearest distributional neighbors of the English adjective "monumental" (cited by [8]).

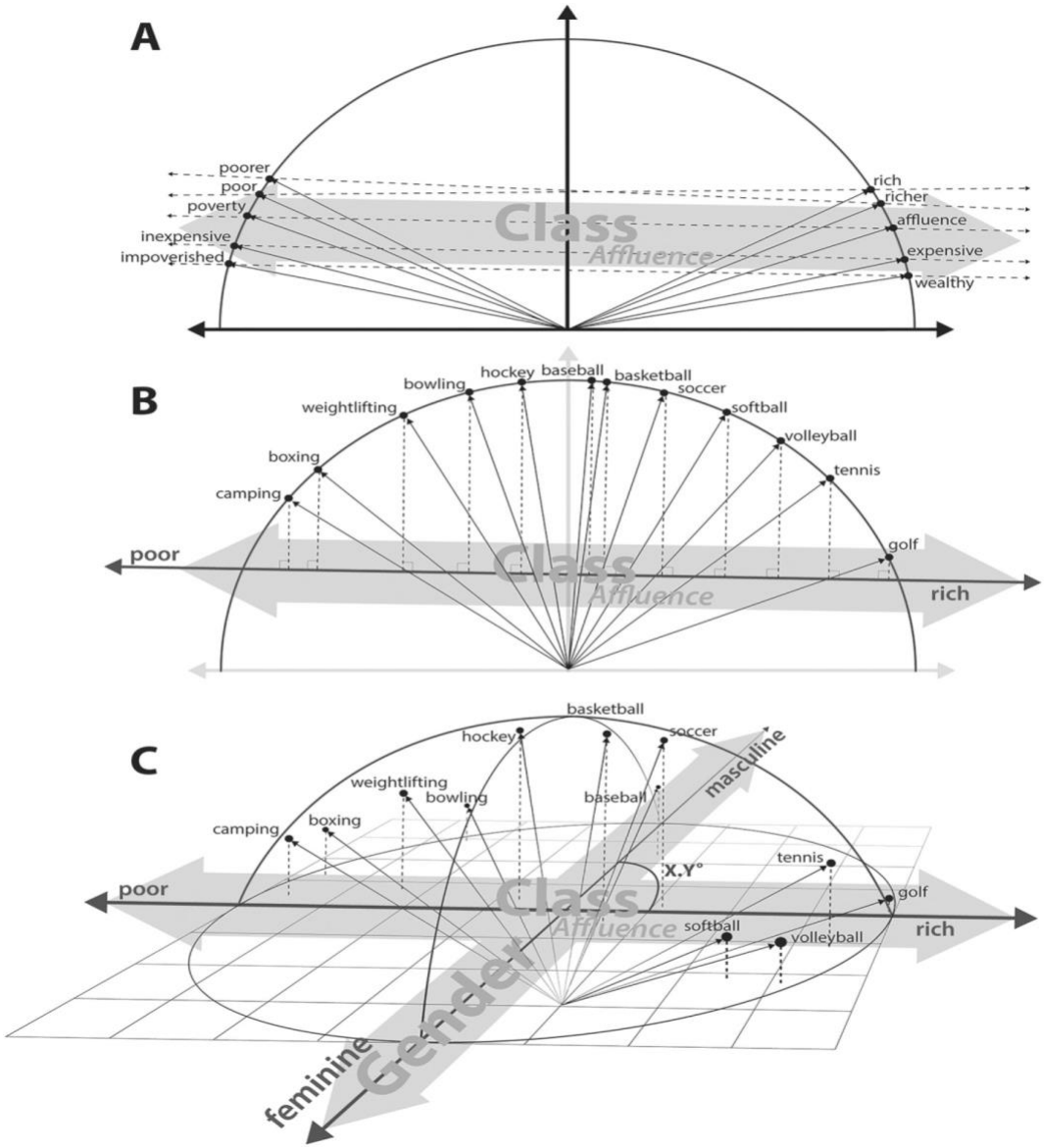

Figure 5 – Conceptual Diagram of (A) the Construction of a Cultural Dimension; (B) the Projection of Words onto That Dimension; and (C) the Simultaneous Projection of Words onto Multiple Dimensions (cited by [9])

Large topic groups can be extracted using the following algorithm. Let us have a dictionary $\mathbf{d} = \{w_i\}$. Using Word2Vec model as a source for vectors, we can shape a matrix $\mathbf{E} = \text{Word2Vec}(\mathbf{d})$ consisting of vectors for every word from the dictionary. Let n be a counter of passed steps, let $n = 1$ and $\mathbf{w} = \mathbf{d}$. Thus, using matrix $\mathbf{E}$ and dictionary $\mathbf{d}$, we can introduce the following algorithm of word separation.

1. Calculate the matrices $\mathbf{E} = \text{Word2Vec}(\mathbf{w})$, $\mathbf{R} = \text{LSA}(\mathbf{E})$ – an ordered set of axes in a reduced space.
2. Take n^{th} vector from matrix $\mathbf{R}$: $\mathbf{r} = \mathbf{R}_n$, consider the vector $\mathbf{r}$ as an axis in the latent embedding space.
3. Sort all words according to their values along the axis $\mathbf{r}$: $\mathbf{w'} = argsort(\mathbf{w}, \mathbf{r})$.
4. Let us divide the sorted list of words $\mathbf{w'}$ into three equal sub-lists according to their values by axis $\mathbf{r}$: $\mathbf{d'} = <\mathbf{d}^-, \mathbf{d}^o, \mathbf{d}^+>$. Let $n = n + 1$. Until we made a given number of algorithm's steps, repeat the algorithm for dictionaries $\mathbf{d}^-$ and $\mathbf{d}^+$.

The result of this algorithm should be a hierarchy of axes (vectors) providing separation of a vector space into semantically related parts. Note that such a hierarchy presented by a tree with more common and abstract properties closer to the tree's root. The tree-shaped structure of the new space looks reasonable according to the common sense: two words belonging to two different classes could not have common properties in case of these classes do not have a third class in common. That is why we are separating words into different classes – they are different because they do not share the same properties; e.g., abstract concepts could not have dimensionality and other features of physical objects.

Note that at every step we select two sub-dictionaries, $\mathbf{d}^-$ and $\mathbf{d}^+$, and eliminate one of them, $\mathbf{d}^o$; that is why we construct a binary tree of sub-dictionaries.

For the sake of interpretation of results, we used the Russian Wictionary to compose topic word lists for the following domains: geology, geological epochs, geography, minerals, plants, weapon, arts, philology, philosophy, informatics, architecture, fortification, politics, names of professions, military and civil ranks, male and female names, old lexis, Russian cities and rivers, animate nouns. We used these categories for visual evaluation of resulting separation of dictionaries. We paid a special attention to select topics which are far from the most of other topics but have one or two in neighbor for the sake of visual separability of results – e.g, old lexis vs modern one, humanity vs natural science etc. Usage of the same topic lists allows comparison of representation at different layers of resulted hierarchy.

Our method for visual analysis of results is the following. We draw a heat-map for every layer of separation. A heat-map here is a table which cells are colored using a gradient palette; a column in this table corresponds a sub-dictionary at a selected layer of hierarchy, a row in this table represents one of the topic lists. For every row we calculated the distribution of words from topic list among sub-dictionaries of the given layer (for every cell we used intersection of two lists). For the sake of better visual interpretation we used the formula (2) for a value v_{ij} in a sub-dictionary number i and topic list number j.

$$v_{ij} = \frac{log(1+words_{ij})}{1+words_j}. \quad (2)$$

Here $words_{ij}$ – the number of words in intersection of sub-dictionary i and topic list j, $words_j$ – number of words in topic list j. This normalization rises contrast ration of the resulting image; we need such normalization since usage of linear normalization decreases specificity of the whole picture.

One heat-map demonstrates separation of dictionaries for just one layer; thus, we need several images to represent the whole hierarchy. Note that we cannot control the direction of axes calculated by LSA. This means that moving from one layer to the next one we cannot guaranty that the order of sub-dictionaries will be the same. Thus, this means that far sub-dictionaries in the embedding space could become neighbor columns in the heat-map.

4. Visual analysis of word separation at the top levels of the hierarchy

This section demonstrates results of our experiments with a Word2Vec model trained on scientific articles written in architecture, arts, automatization, geology, history, linguistics, and literature. We also used pre-trained models from site RusVectōres; however, their analysis is beyond scope of this article. Note that results for these last models demonstrated the same quality of visualization.

Figure 6 demonstrates the separation after the first step of separation. It is easy to see that words belonging to geology, names of minerals and cities are constituting the same group. This group also contains some words from philology, phylosophy, politics, and some of animated nouns.

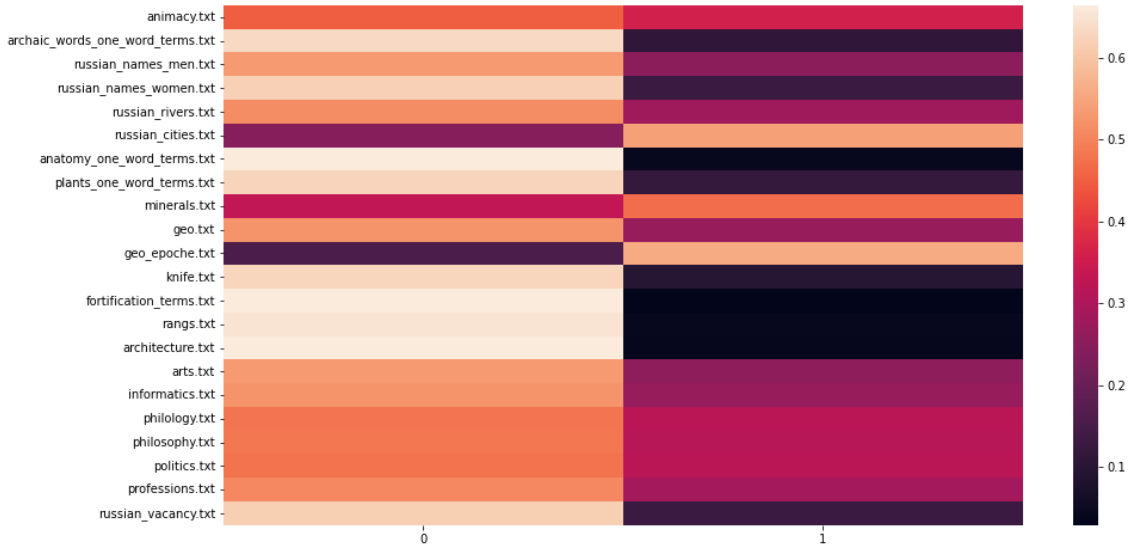

Figure 6 – Separation of words at the first layer of hierarchy

The more detailed analysis of results demonstrates that the first layer of achieved hierarchy opposes a colloquial vocabulary to a scientific one. The words with maximal values along the first axis are belonging to everyday conversations, the words with minimal values consist of surnames of scientists and authors of articles, names of universities and research organizations, cities, special terms, e.g. *intertextuality* and *linearization*. Despite the fact that the difference between two sub-dictionaries can be described as the difference between colloquial and scientific discourses, there are scientific terms in both parts. However, the semantic complexity of terms from "scientific" part of the space is higher than from the "colloquial" one. For example, the term *poem* belongs to "colloquial" while terms *accentual verse* and *acrostic* belong to "scientific". The same is true for informatics; words *register*, *code*, *argument*, *mail*, *core*, *module*, *computer*, *buffer*, *scenario*, *container*, *subject*, and *protocol* were placed at an opposite side of the axis with such words as *cashing*, *compiler*, *replication*, *octet*, *encapsulation*, *coding*, *tracker*, *emulation*, *bit rate*, *profiling*, *quantifier*, and *selector*. Note that the former words have a higher probability of occurrence in a news wire or small talk than the later ones while the later words occur more often in scientific papers or manuals.

Figure 7 demonstrates that sub-dictionary at the third layer have a closer context. Sub-dictionaries are containing words belonging to the following topics: (0) names of professions, names of ranks, politics; (1) philosophy and philology; (2) names of plants, minerals, weapons, geographic, fortification, and architectural terms (those are more frequently occurred in "scientific" discourse than in "colloquial" one); (6) names and surnames; (7) cities and organizations. Sub-dictionaries (4) and (5) contain a few of words because of their specificity. It is easy to see that names of researchers and organizations belong to "scientific"

discourse. The resulting hierarchy of separation constructed at the third layer is presented inTable 1.

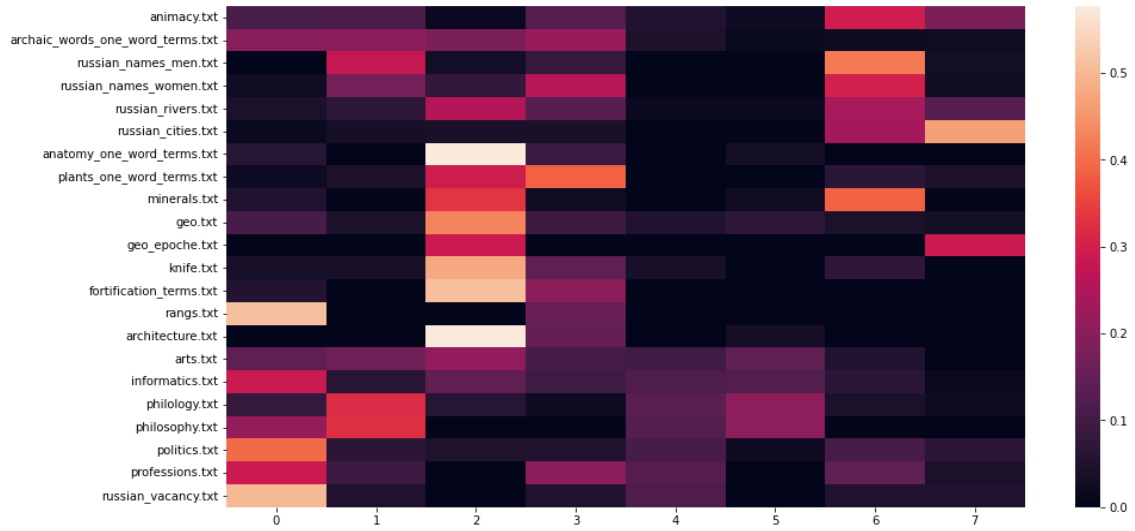

Figure 7 – Separation of words at the third layer of hierarchy

Table 1 – Topics at the third layer of hierarchy

colloquial discourse				scientific discourse			
abstract		physical		scientific terms		places and people	
society and politics	books and religion	special terms	Everyday items	inner scientific	common scientific	names of researchers	places and organizations

Our first hypothesis was that the first layer divides the model's vocabulary into different scientific areas; however, our experiments demonstrated that division at first several layers uses more abstract and universal features. We found out that top layers of the investigated vector space devoted to abstract lexis which is common to every area and is used for description of the same ideas: task statement, introductory phrases, method description etc. Specific lexis of different scientific areas could be found at more deeper layers. Scientific terms, placed at the second language, is divided into specializations, inner scientific processes (*investigation, improvement, specialization, acquisition, studying, thesis, methodology*), and phenomena, which are constituting one pole of the axis, and properties and processes carried out over objects of science (*analyticity, stereotypicy, subjectivity, precedence, obligingness, dissociation, asymmetricity, locality*).

At the 6th layer, we extracted groups consisting of about 30 words (see Figure 8). Our algorithm reflects the overall tendency, but the probability of co-occurrence a word from a sub-dictionary in a topic list is very small. Thus, our method of visual representation of results suffers crucial drawbacks. Since we do not filter results by their frequencies, such categories as name of ranks and geology consist merely one term, which demonstrates maximal normalized value on the heat-map.

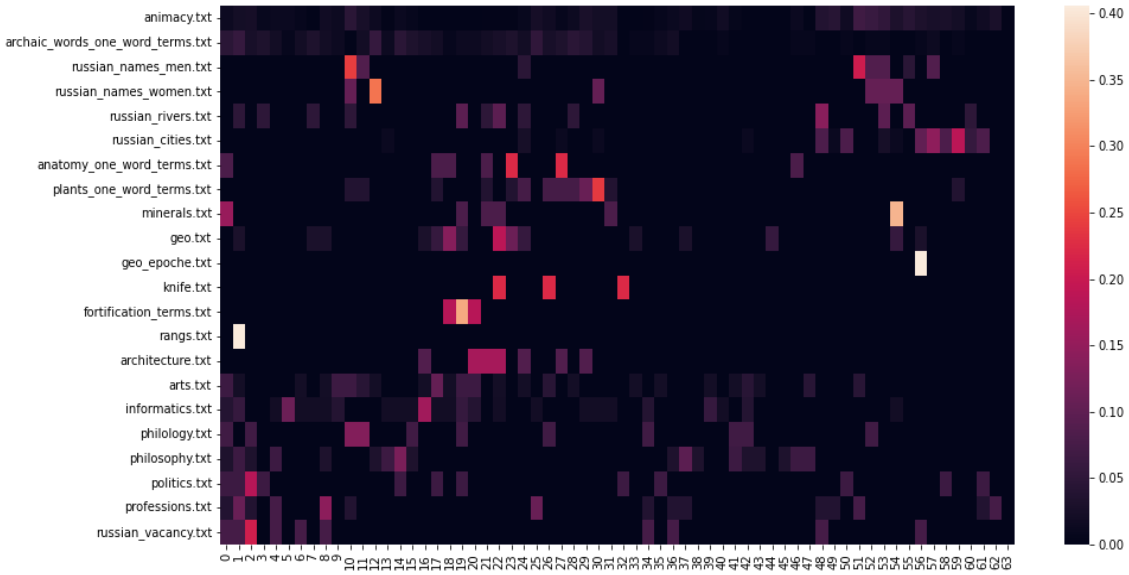

Figure 8 – Separation of words at the sixth layer of hierarchy

As it was mentioned above, we applied our method to other vector embedding models. Interpretation of these model differs from the described above. However, all of these models share some common features at several top layers. These are separation at colloquial and special lexis, abstract and physical, material and ideal. Note that the same feature can be found at different layers for different models, thus, there is no one universal principle for shaping of the features hierarchy.

6. Conclusion

In this paper we proved a hypothesis that Word2vec vector embedding spaces can be split into interpretable groups not only at local level, but also for the model as a whole. Our investigation demonstrated that the logic of such partitioning depends on style and source of texts used for training of a model. For example, a model trained on belletristic of different epochs separates abstract and concrete at the first layer, opposes moral to colloquial and modern to archaic at the second layer. A model trained on Internet texts is divided into social and organizational at the first layer, and opposes abstract to concrete and technology to control at the second layer.

Our results were analyzed using a method of visual representation of distribution of topic lists among sub-dictionaries; the method based on a heat-map demonstrating a share of words belonging both to a topic list and a sub-dictionary. The method provides a good visualization; however, it suffers some drawbacks. For example, at deep layers of the built hierarchy, the number of words falls exponentially; as result, the probability of finding word from a topic list in a small area of vector space also falls. Moreover, the selected terms happened to be polysemous and belonged simultaneously to several sub-spaces. Note that stricter choice of words for such topic lists needs more linguistic attention.

Analysis of several Word2vec vector models trained on texts of different styles and domains demonstrated that a resulting hierarchy of axis depend on the lexis used in such texts. Some of those axes were selected for every model but at different layers of hierarchy. Thus, we can state that there are some universal axes but not their mutual positions in the resulting space.

Another problem here is selection of borders between sub-dictionaries. As it was mentioned above, we merely divided a dictionary into three sub-dictionaries having the equal number of words according to their coordinates along calculated axes. Such a border could separate words from the same semantic group and decreases accuracy of our method. We hope that preliminary clustering of words belonging to borderline could solve this problem.

Finally, we analyzed just words on periphery, while most words are composing a dense cluster in the center. However, the density of such clusters prevents their correct separations. Thus, this problem needs a special investigation.

6. References

[1] Mikolov T., Chen K., Corrado G., Dean J. Distributed Representations of Words and Phrases and their Compositionality // In Proc. of Neural Information Processing Systems 27: 27th Annual Conference on Neural Information Processing Systems, 2013. P.3111-3119

[2] Korogodina O., Karpik O., Klyshinsky E. 2020. Evaluation of Vector Transformations for Russian Word2Vec and FastText Embeddings // Conference on Computer Graphics and Machine Vision: GraphiCon 2020. 2020. V. 2744. P. paper18-1 – paper18-12.

[3] Korogodina O., Koulichenko V., Karpik O., Klyshinsky E. 2021. Evaluation of Vector Transformations for Russian Static and Contextualized Embeddings // Conference on Computer Graphics and Machine Vision: GraphiCon 2021. 2021. V. 3027. P. 349-357.

[4] B. Wang, A. Wang, F. Chen, Y. Wang, J. Kou, Evaluating word embedding models: methods and experimental results // In Proc. of APSIPA Transactions on Signal and Information Processing, 2019, 8. doi: 10.1017/ATSIP.2019.12

[5] Lasri K., Pimentel T., Lenci A., Poibeau T., Cotterell R. Probing for the Usage of Grammatical Number / In Proc. of the 60th Annual Meeting of the Association for Computational Linguistics, V. 1, pp. 8818–8831.

[6] Conneau A. et al. What you can cram into a single vector: Probing sentence embeddings for linguistic properties [Электронный ресурс]: arXiv preprint arXiv:1805.01070. – 2018. URL: https://arxiv.org/abs/1805.01070 (дата обращения 01.10.2022).

[7] Ravfogel S. et al. Counterfactual interventions reveal the causal effect of relative clause representations on agreement prediction [Электронный ресурс]: arXiv preprint arXiv:2105.06965. – 2021. URL: https://arxiv.org/abs/2105.06965 (дата обращения 01.10.2022).

[8] Kutuzov A. Distributional word embeddings in modeling diachronic semantic change // Doctoral Thesis, University of Oslo, [Электронный ресурс]: https://www.duo.uio.no/bitstream/handle/10852/81045/1/Kutuzov-Thesis.pdf (дата обращения 01.10.2022).

[9] Kozlowski A., Taddy M., Evansa J. The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings. American Sociological Review, 2017, pp. 905-949.

[10] Воронцов К. В., Потапенко А. А. Регуляризация вероятностных тематических моделей для повышения интерпретируемости и определения числа тем // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». — Вып. 13 (20). — М: Изд-во РГГУ, 2014. — С. 676-687.

[11] Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K., Harshman R. Indexing by latent semantic analysis // Journal of the American Society for Information Science. 1990. V. 41, Iss. 6. P. 391-407.