An approach to visual thesaurus exploration: a case study for Russian language

Stanislav Protasov, email: s.protasov@innopolis.ru

Innopolis University

Аннотация. *Big problem visualization at a single plot can be considered as a handy tool for making managerial decisions. Thesaurus can be used to represent important properties of a large text corpus. In this work I propose fully automatic method of visualizing text dataset thesaurus on a plain graph which can be used for domain exploration. Resulting visualization shows quality semantic clusters and it is very useful for identifying items and clusters underrepresented in the dataset.*

Ключевые слова: *knowledge representation, visualization, dimensionality reduction, natural language processing*

1. Introduction

Making a management decision often requires having a top view on a problem to see it as a whole. Complex problems may include multiple components with no simple and obvious visual form; thus, visualization can become a difficult task itself. For example, topographic objects can be easily represented as dots and lines on a map, physical experiments can be visualized with a histogram, computer networks can be plotted as a graph. On the other hand, collection of images, texts of audio files cannot be easily mapped on a single image. For a single object it is very important to have a universal low-dimensional representation of essential features; multimedia data does not have such default representation.

Another important thing that one needs to consider while visualizing big datasets, is that frequently not all information for decision making is contained within the dataset. For example, “far” and “close” relations will depend on the context you put the data in: whether this is a map of a city or world map. Another example is putting data on a height map. Even if two items have very close coordinates in a dataset, one can be located in a basement while the other is on the 100th floor of a skyscraper.

In this work I target both abovementioned problems applied to a big *thesaurus* visualization. Here I understand thesaurus as a vocabulary of the most representative words in a corpus. Main work contribution is an approach to prepare a thesaurus for a big text collection and visualize it in an arbitrary context for the task of exploration.

The work is organized as follows. In the *Related works* section I overview existing thesaurus visualization approaches and methods I use for

my solution. In the *Methodology* section I construct the whole visualization pipeline. *Results* and *Discussion* sections are devoted to results, problems, and observations. In the Conclusion part I summarize results of the work.

2. Related works

In this section I will discuss two major directions. Firstly, I will cover existing approaches to low-dimensional visualization of high-dimensional data. Secondly, I will overview approaches important specifically for natural language thesaurus visualization.

To visualize any high-dimensional data one will need to introduce a method which extracts essential information about the data items of the dataset as 2 or 3-dimensional vectors. There exist multiple methods to obtain low-dimensional approximation. Ones will try to find general tendencies in the data, highlighting outliers, but as these methods are robust to small deviations, they can fail to correctly project item minorities. Still, robustness of such methods allows to prepare static embedding models which will be actual even if dataset insignificantly changes with time. Such methods are well-studied and include, for example, principal component analysis [3] which tries to minimize global variance loss for the dataset, random projections [11] which targets to preserve Euclidean distances in a new low-dimensional metric space and pivot-mapping technique [12] which is used in data indexing and tends to provide false-positive similarities. Other methods pay more attention to local clusters of items and can be used to highlight closeness of items belonging to the same cluster but will fail to correctly project large distances. Among them are elastic maps [1], self-organizing maps [2] and t-SNE [4] widely used in practice, as in [7] for visualizing medical reports. All these methods are often used in clustering pipelines as they preserve dense clusters. Thus, I will suggest using global methods for overview tasks and local methods for exploration.

Thesaurus visualization assumes that extracted collection of terms (words or phrases) from a text corpus is then firstly represented in numerical form. For numerical representation or terms collection there are two basic approaches: graph-based and vector-based. The most well-known graph-based model is WordNet [13] project, which provides a powerful tool to explore different word relations. Major weakness of this and similar models is that it must be maintained. On the other hand, vector-based methods exploit distributional hypothesis [14] which allows to obtain latent space embeddings for words, sentences, and texts in an unsupervised manner. This idea survives since latent semantic analysis [15], it got a significant boost in application with word2vec models [9] and today it is exploiting attention mechanism [8] to build human-level conversation, translation, question-answering and other natural language processing solutions. Big variety of

embedding models gives a researcher or an engineer a wide choice or features including speed, accuracy, memory consumption, robustness, and their combinations. For example, word2vec [9] and GloVe [6] models are extremely fast and compact, BERT [8] and fastText [10] solve the problem of typos and out-of-vocabulary tokens, pretrained compact word embeddings like those in Navec (github.com/natasha/navec), RusVectores (rusvectores.org/ru/models) allow constant-time inference. All these models and methods either build a high-dimensional representation for a dataset from scratch or allows to reuse models trained for similar domains.

It is important to mention that a context also can be a problem for thesaurus visualization. Any special text corpus always has biased distributions in a language model. Both global and local low-dimensional approximators are very sensitive to context change, thus one should have a clear understanding of a context to use as a "background". One of the most universal sources of rich lexicon is Wikipedia dumps. One can use the whole corpus, while in my work I utilize page titles of the main namespace to represent scientific background. Other good lexicon sources are the national language corpora as the Corpus of Contemporary American English (COCA) [15] or National Corpus of Russian Language (NCRL) [5].

3. Methodology

In this work I propose a full pipeline of extracting, preparing and visualizing thesaurus. I must note that in my experiments I used a private dataset of popular science texts in Russian, but one can easily replace it with a similar popular science text collection (e.g. *nts-lib.ru* or similar) to reproduce results.

Thesaurus extraction. In this phase I suggest ensuring which characteristics you consider the most important for representative words. In this work I tried two extraction techniques. One technique is based on named entity recognition (NER) methods from libraries like Natasha (github.com/natasha) and SpaCy (spacy.io). This method builds a quality thesaurus of names, organizations, and locations. Unfortunately, these words usually fall out of vocabulary, thus embedding quality for them can be very low. Another thesaurus extraction approach is to use language models to find discrepancies between dataset word distributions and distributions in some reference corpus. I used unigram distribution models of a dataset and NCRL model obtained from official website. For discrepancy detection I used Kullback-Leibler divergence formula for a single summand:

$$D(term) = P_{dataset}(term) * \log\left(\frac{P_{NCRL}(term)}{P_{dataset}(term)}\right)$$

To be accepted as a part of thesaurus, the term should satisfy empirically chosen thresholds. In my experiments these are $P_{NCRL}(term) > 0.005$ and $D(term) < -0.015$. Selected words are then merged with the result of NER.

Thesaurus exploration task assumes that we are not only interested in what is present in the data, but also what is missing. Thus, there should be a background context dataset which can be compared with a contour map for data visualization. In my experiments I used page titles of the Russian Wikipedia main namespace.

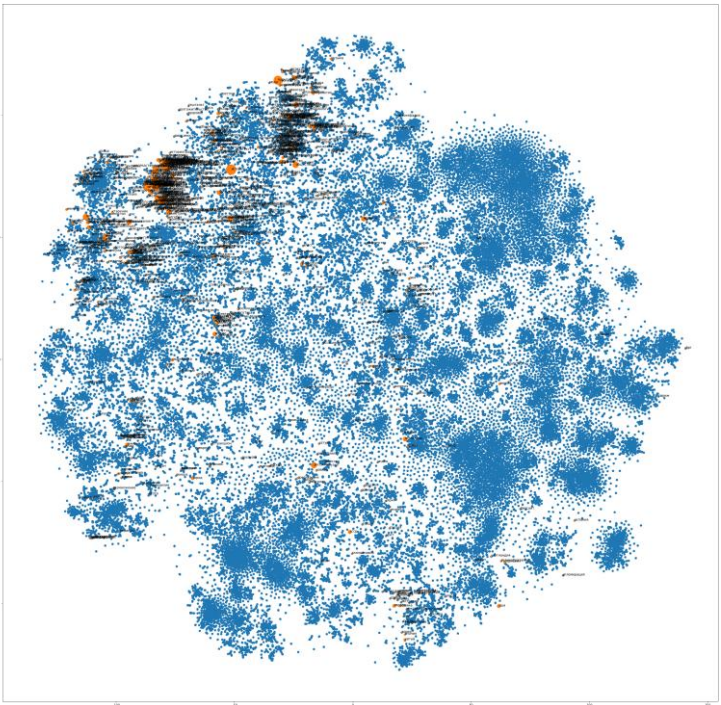

Figure 1. t-SNE mapped dataset thesaurus (orange) visualization on a background of 70000 Russian Wikipedia titles (blue)

Vector representation. Next stage is obtaining vector representation for thesaurus and background words. In my experiments I used GloVe embeddings from Navec project and fastText embeddings from RusVectores precomputed for a large lexicon of Russian language. Deeper networks can be used as in [7] if visualization focus falls more on semantics and syntax, rather than particular words. The choice of pretrained models is essential in

my case, as for chosen context corpus has 70000 known words, thus inference time can be very big.

2D mapping. After obtaining vectors for both context (Wikipedia) and a dataset, I merge them to create a common term collection. For this collection I run dimensionality reduction procedure. As it was mentioned in the *Related works* section, local methods like t-SNE tend to preserve dense clusters. This is a highly important feature, as the exploration task can include both identifying missing terms in semantic clusters and underrepresented clusters. Thus, I used t-SNE algorithms with perplexity value equals 40 and 3000 iterations.

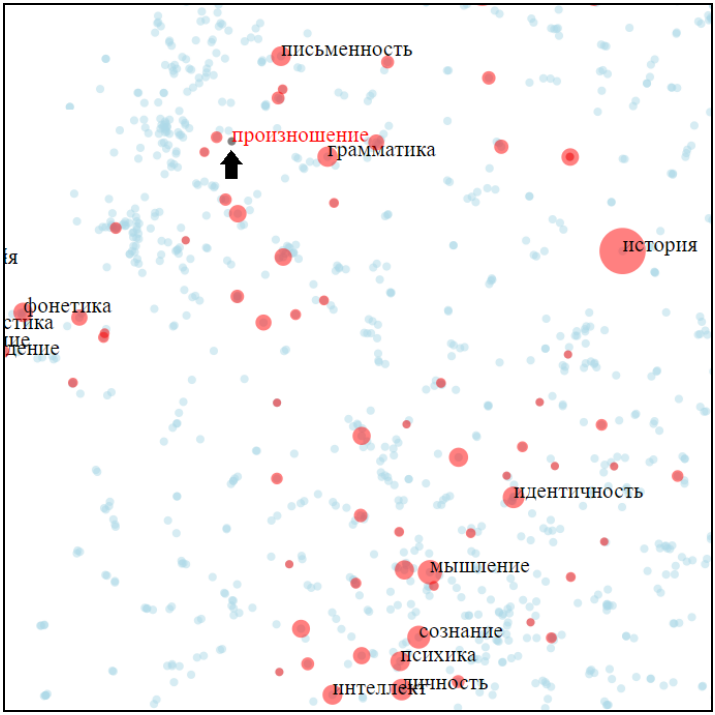

Figure 2. Sample of the generated SVG map with the highlighted term “произношение” on mouse hover.

Visualization. Last step in visualization is to prepare a compact, portable, and server-independent solution. There are few things that are important to consider while building and image. Firstly, representing words without labels is useless, while displaying all 70000 labels will overload

visualization; thus, I print out only words with the highest frequency in a dataset and show other labels on hover. Secondly, if there is a term frequency information, it can be encoded with marker area. To fulfill these requirements, I chose to generate an HTML file with SVG image included. Markers are <*circle*> tags, and *:hover* CSS pseudoclass is used to display annotation on hover over the marker. Background markers are blue and of the same size. Dataset marker area size is chosen with respect to the frequency in a dataset. Such configuration allows to build a relatively small (15 MB) HTML file which works without javascript support in all major browsers even on smartphones.

Resulting map overview is presented at figure 1. Sample part of the map is shown at figure 2.

4. Results

Whole processing visualization was done on a 2-core Intel Core i5 CPU. The most time-consuming stages are thesaurus extraction with a language model and t-SNE algorithm. First took 7 CPU hours to obtain lemmatized tokens from 96 MB of UTF-8 text with pymystem library. Second converges in half an hour for 70000 items represented with 300-dimensional vectors. Other stages are relatively short, rendering of HTML file takes up to 5 minutes, displaying prepared HTML document is fast in Mozilla Firefox and can take up to 2 minutes in Chrome. Loaded document occupies near 500 MB of RAM in Mozilla Firefox browser.

5. Discussion

Provided method gives a good starting point to overview the domain of some text dataset. Firstly, as it can be seen in figure 1, t-SNE method prepares quality semantic clusters, which can be easily labelled. I was able to identify “war”, “Formula 1 racers”, “name of sciences” clusters and so on. Secondly, visualization indeed helps to find poorly covered clusters, missing topics in densely covered clusters, which can be directly used for making management decisions.

The method can be criticized for the fact, that there is no clear pattern of similarity. Some clusters are formed by the common topics (e.g. war), others consist of surnames (e.g. “F1 racers” or “chess players”), some are glued by both semantics and grammar (e.g. “names of sciences” or “names of drugs”). Thus, even if there is a clear pattern inside the cluster, it changes as soon as you cross its border. Global methods like PCA can provide a common pattern for the whole dataset, but they are worse in visualizing dedicated clusters.

One more observation is that some Wikipedia page names are irrelevant for thesaurus context. In the resulting map I was able to find sparse regions

with German surnames of just “first names” areas. These titles can be filtered out on a preprocessing stage.

6. Conclusion

In this paper I showed an approach to visualizing an arbitrary dataset thesaurus using a context of a big lexicon (Wikipedia). Proposed method is fully automatic and generates a visualization for 96 MB dataset in Russian in 7 hours. Resulting plot is saved as an HTML file and can be opened in all major browsers, including smartphones. Resulting visualization shows quality semantic clusters and it is very useful for identifying gaps of small (single term) and large (unconsidered clusters) size.

Proposed method has some disadvantages, related to the fact that similarity within clusters has different nature. This issue can be addressed with replacing the t-SNE method with a global dimensionality reduction approach like PCA.

In general, I was able to propose a handy portable interactive visualization, which can be used to make management decisions related to content of the dataset.

Bibliography

1. A. N. Gorban A. N., Zinovyev A.Y. Principal Graphs and Manifolds //Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques: Algorithms, Methods, and Techniques. – IGI Global, 2009.
2. Kohonen T. Self-organized formation of topologically correct feature maps //Biological cybernetics. – 1982. – Т. 43. – №. 1. – С. 59-69.
3. Pearson K. LIII. On lines and planes of closest fit to systems of points in space //The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science. – 1901. – Т. 2. – №. 11. – С. 559-572.
4. Hinton G. E., Roweis S. Stochastic neighbor embedding //Advances in neural information processing systems. – 2002. – Т. 15. – С. 857-864.
5. Апресян Ю. Д. и др. Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы //Национальный корпус русского языка. – 2003. – Т. 2005. – С. 193-214.
6. Pennington J., Socher R., Manning C. D. Glove: Global vectors for word representation //Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). – 2014. – С. 1532-1543.
7. Kivotova E. et al. Extracting clinical information from chest x-ray reports: A case study for Russian language //2020 International Conference Nonlinearity, Information and Robotics (NIR). – IEEE, 2020. – С. 1-6.

8. Vaswani A. et al. Attention is all you need //Advances in neural information processing systems. – 2017. – С. 5998-6008.

9. Mikolov T. et al. Efficient estimation of word representations in vector space //arXiv preprint arXiv:1301.3781. – 2013.

10. Bojanowski P. et al. Enriching word vectors with subword information //Transactions of the Association for Computational Linguistics. – 2017. – Т. 5. – С. 135-146.

11. Yang K. et al. Distributed Similarity Queries in Metric Spaces //Data Science and Engineering. – 2019. – Т. 4. – №. 2. – С. 93-108.

12. Bingham E., Mannila H. Random projection in dimensionality reduction: applications to image and text data //Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining. – 2001. – С. 245-250.

13. Miller G. A. et al. Introduction to WordNet: An on-line lexical database //International journal of lexicography. – 1990. – Т. 3. – №. 4. – С. 235-244.

14. Harris Z. S. Distributional structure //Word. – 1954. – Т. 10. – №. 2-3. – С. 146-162.

15. Deerwester S. et al. Improving information-retrieval with latent semantic indexing //Proceedings of the ASIS annual meeting. – 143 OLD MARLTON PIKE, MEDFORD, NJ 08055-8750 : INFORMATION TODAY INC, 1988. – Т. 25. – С. 36-40.

16. Davies M.. The Corpus of Contemporary American English (COCA) [Available online] : 2002. — https://www.english-corpora.org/coca/.