

Can Eye Movement Data Be Used As Ground Truth For Word Embeddings Evaluation?

Amir Bakarov

The National Research University Higher School of Economics, Moscow, Russia
Federal Research Center ‘Computer Science and Control’ of Russian Academy of Sciences, Moscow, Russia
amirbakarov@gmail.com

Abstract

In recent years a certain success in the task of modeling lexical semantics was obtained with distributional semantic models. Nevertheless, the scientific community is still unaware what is the most reliable evaluation method for these models. Some researchers argue that the only possible gold standard could be obtained from neuro-cognitive resources that store information about human cognition. One of such resources is eye movement data on silent reading. The goal of this work is to test the hypothesis of whether such data could be used to evaluate distributional semantic models on different languages. We propose experiments with English and Russian eye movement datasets (Provo Corpus, GECO and Russian Sentence Corpus), word vectors (Skip-Gram models trained on national corpora and Web corpora) and word similarity datasets of Russian and English assessed by humans in order to find the existence of correlation between embeddings and eye movement data and test the hypothesis that this correlation is language independent. As a result, we found that the validity of the hypothesis being tested could be questioned.

Keywords: word embeddings, eye-tracking, distributional semantics, evaluation

1. Introduction

Dense vector word representations (*word embeddings*) are gaining popularity in the scientific community because of their proved efficiency in certain downstream task. However, the we are still unaware what is the most reliable evaluation method for these models. Construction of a dataset for certain downstream tasks is expensive, and some researchers argue that it is better to have a universal dataset on which can be evaluated the ability of task independent word vectors to model the structure and space of lexical semantics (or different aspects of semantics) (Schnabel et al., 2015).

Evaluation on such data is usually called *intrinsic evaluation* (as an opposite to the *extrinsic evaluation* which is an evaluation on downstream tasks). However, it is not actually clear how to evaluate modeling of lexical semantics. Different intrinsic evaluation tasks propose different notations of semantics, but most of the existing methods of intrinsic evaluation (like *word similarity* method or *word analogy* method) are criticized (Batchkarov et al., 2016; Rogers et al., 2017). Thus, some researchers started to propose experimental evaluation techniques trying to use *neuro-cognitive resources* as a gold standard for semantic modeling.

Among these techniques, there are methods of detecting neural activation patterns on text processing (like *magnetoencephalography* (Wehbe et al., 2014)), methods of detecting reaction time on reading (like *semantic priming* (Auguste et al., 2017)), and many others. In this work, we propose another possible data source which is *eye movement on silent reading*. This data consists of measurements of time of gaze fixation on each word, amount of returns of the gaze, and so on. Eye movement data is supposed to be related to human language processing and possibly stores information about lexical semantics, so hypothetically the correlation with word embeddings could be found. Similar

experiments with English eye movement data were already proposed in (Søgaard, 2016), and the results did not show strong correlation. However, it is not clear whether the same outcome will be observed in other languages. If eye movement data is language dependent, that this could support the idea that those data are less related to the semantics. So, the *main aim of this paper* is to answer the question whether word embeddings in one language correlate with eye movement data in another language (and how much). Of course, we cannot talk about investigating language independence (for that we would need many more typologically diverse languages), but possibly we could highlight some features of this comparison investigation two distinct languages.

This study proposes experiments with eye movement data and word embeddings for Russian as ‘another’ language (we consider Russian because we are native speakers of this language, so we are able to interpret the obtained results), trying to answer the question of whether word embeddings in one language correlate with eye movement data in another, and whether the correlations between two language specific embeddings and eye movement data are comparable. Probably, distributional semantics and eye movement data process different types of semantics, so we could expect that the correlation would be low.

This paper is organized as follows. Section 2 puts our work in the context of previous studies. Section 3 describes the data that we are using in the current work. Section 4 is about the experimental setup. The results of the comparison are reported in Section 5, while Section 6 concludes the paper.

2. Related work

The basic idea of this work is related to the hypothesis that observable features of human text processing (like **the time of reading of a certain word**) are based not only on the surface features of a linguistic sign but also on its meaning.

Target word	The nearest neighbor word		
	RSC, <i>gaze</i>	Araneum, <i>embeddings</i>	Ruscorpora, <i>embeddings</i>
mikrob (<i>microbe</i>)	lekarstvo (<i>cure</i>)	boleznetvorniy (<i>pathogenic</i>)	mikroorganizm (<i>microorganism</i>)
speciya (<i>spice</i>)	kuriniy (<i>chicken</i>)	priprava (<i>seasoning</i>)	speciya (<i>spice</i>)
zanyatie (<i>class</i>)	student (<i>student</i>)	trenirovka (<i>training</i>)	klassniy (<i>in class</i>)
pluvat (<i>to cheat</i>)	ministr (<i>minister</i>)	tyrit' (<i>to steal</i>)	vorovat (<i>to steal</i>)
chaska (<i>cup</i>)	vedro (<i>bucket</i>)	stakan (<i>glass</i>)	chaynik (<i>kettle</i>)

Table 1: Top-1 nearest neighbors for Russian gaze vectors (first column) and Russian embeddings vectors (second and third columns). The first word is transliteration in Russian, the word in brackets is translation into English.

There are certain studies based on eye movement data that prove correlation between word semantics and reading time (Smith and Levy, 2013; Hohenstein, 2013). More precisely, while reading a human’s brain continuously builds a model of context for already read words and integrates each new word in context, comparing it with contexts stored in the memory. The effort of this integration is inversely proportional to how probable the word is, so when an encountered word is highly unpredictable, then the time of its reading should increase.

This research is strongly inspired by (Søgaard, 2016), which evaluated word embeddings against eye movement data from the Dundee Corpus (Kennedy et al., 2003) using aggregate statistics of eye movement data features. Their experiments showed that there is no notable agreement between eye movement data and word embeddings. However, these experiments were performed for English only. In our work we extend the propose of (Søgaard, 2016), making similar experiments for English and Russian to test the hypothesis whether the same outcome will be observed in other languages.

3. Datasets

3.1. Eye movement data

Eye movement data is obtained through *eye-tracking*, the process of measuring the point of the human gaze on the screen. In other words, when a person reads text on the screen, a special mechanism called eye-tracker tracks the movement of the gaze and records the information about the reading. A number of different features can be recorded, e.g. how long the gaze was fixated on a certain word, how many times the gaze returned back, and so on. We averaged the data of all examinees to obtain a feature vector for each words, so in these vectors each component would report value of each of the tracked features. We normalized these values, obtaining values of each of the features ranging from -1 to 1, and used vectors of normalized eye movement features as a gold standard for word embeddings evaluation (here and later we will us the notion of **gaze vectors**).

Russian. As a source of Russian eye movement data we used *Russian Sentence Corpus (RSC)* (Laurinavichyute et al., 2017) which contains data about reading 144 Russian sentences by 96 native speakers. After averaging examinees scores for each token and averaging word form scores for each lemma, we obtained a dataset with information on eye

movements for 701 single words.

English. We used an English eye movement corpus which is the *Provo Corpus* (Luke and Christianson, 2017) (the *Dundee Corpus* used in (Søgaard, 2016) is not publicly available). It contains data on reading 55 English paragraphs by 84 native speakers. We obtained vectors with information on eye movements for 1185 words from this data (with the manipulations described above). We are also aware of another publicly available English corpus, *Ghent Eye-Tracking Corpus (GECO)* (Cop et al., 2017) containing data on reading 5 000 English sentences by native and bilingual (for which English is a second language) English speakers (33 participants overall). In this paper the data of native speakers only is used. We obtained gaze vectors for 987 words.

In all cases the raw data consisted of 17 logical and continuous eye movement features. The features included:

1. *dwelt time* (summation of the duration across all fixations) on the current interest area;
2. duration of the first fixation event that was within the current interest area;
3. dwell time of the first run within the current interest area;
4. number of all fixations in a trial falling in the first run of the current interest area;
5. total fixations falling in the interest area;
6. whether the first fixation in the interest area N was preceded by a fixation in the interest area $N - 1$;
7. whether the current interest area received at least one regression from the later interest areas;
8. whether regression(s) was/were made from the current interest area to the earlier interest areas;
9. dwell time from when the current interest area is first fixated until the eyes enter an interest area;
10. dwell time of the second run of fixations within the current interest area;
11. no fixation occurred in the first pass reading;
12. the duration of the first fixation made on the interest area $N + 1$ after leaving the interest area N in the first pass;
13. landing position in the word of the incoming saccade;
14. direction of the incoming *saccade* (fast jump from one eye position to another);
15. character that was fixated by the incoming saccade;
16. whether the word was fixated once;
17. whether the word was fixated two or more times.

In all the datasets, the recordings were obtained with an *Eyelink 1000 Plus* desktop mount eye-tracker with a chin rest and a screen on which the sentences were presented.

3.2. Word embeddings

As a distributional model, we use *Skip-Gram*, a neural predictive algorithm that updates values on the input layer (word embeddings), trying to maximize word prediction probability by minimizing the loss of the softmax function (Mikolov et al., 2013). The reason why we employed *Word2Vec* is that it is very common in natural language processing research, evaluated and explored in many papers (note though that (Søgaard, 2016) used another type of embeddings, namely SENNA embeddings (Collobert et al., 2011); however, the paper describing SENNA propose 4 different embedding architectures, and we are not aware which exactly architecture has been used).

Russian. We used a model trained on a POS-tagged *National Russian Corpus* (we will further use the term **Rus-corpora** further) with 195 071 words in the model’s vocabulary (`ruscorpora_upos_skipgram_300_5_2018`) and a model trained on a POS-tagged *Araneum Russicum Maximum (Areneum)* with 196 620 words in the vocabulary (`araneum_upos_skipgram_300_2_2018`) (Kutuzov and Kuzmenko, 2016). The models are available on *Rusvectors* repository¹.

English. We used a model trained on a **Google News** corpus with 2 883 863 words in the vocabulary and a model trained on a POS-tagged British National Corpus (**BNC**) with 163 473 words in the vocabulary (Fares et al., 2017). The models are available on *Nordic Language Processing Laboratory* repository². All used word vectors, both Russian and English, had the same vector dimensionality of 300.

3.3. Word similarity data

Word similarity task is the most ubiquitous technique for word embeddings evaluation. Given words a and b , the task is to find scalar value reporting semantic distance between them. This task is strongly criticized in NLP community, and different researchers address problems like the obscurity of the notion of semantics, subjectivity of human judgments, and so on (Batchkarov et al., 2016).

The word similarity datasets actually differ in the types of human assessments: some datasets are assessed according to semantic similarity relation (which is commonly interpreted as a synonymy, like in words *mug* and *cup*), while other datasets are assessed by semantic relatedness (which is interpreted as co-hyponymy, like in words *cup* and *coffee*).

English. We are aware of more than 7 datasets for word similarity available for English, but in order to propose a fair comparison with gaze vectors we need to drop words that are absent in eye movement data vocabulary. To this end, we did not used datasets like *Verb-143* (Baker et al., 2014), *YP-130* (Yang and Powers, 2006), *RG-65* (Rubenstein and Goodenough, 1965) and *MC-30* (Miller and Charles, 1991), because the amount of remaining word

pairs (after dropping) was too low (lower than 5). We used the only 3 datasets with assessments by semantic similarity.

1. *SimVerb-3500* (234 word pairs remained for GECO, 75 word pairs remained for Provo) (Gerz et al., 2016),
2. *SimLex-999* (37 pairs remained for GECO, 41 pairs remained for Provo) (Hill et al., 2016),
3. *WordSim-353-Similarity* (WS353-Sim) (5 word pairs remained for both GECO and Provo) (Agirre et al., 2009).

We also used 3 English datasets assessed by semantic relatedness, excluding certain datasets for the reasons described above (MTurk-287 (Radinsky et al., 2011)):

1. *MEN* (22 pairs remained for GECO, 77 pairs remained for Provo) (Bruni et al., 2014),
2. *MTurk-771* (7 pairs remained for GECO, 19 pairs remained for Provo) (Halawi et al., 2012),
3. *WordSim-353-Relatedness* (WS353-Rel) (5 pairs remained for GECO, 11 pairs remained for Provo) (Agirre et al., 2009),

Russian. The amount of word similarity datasets available for Russian is lower. All datasets we are aware of are translated versions of English datasets: *SimLex-999*, *WordSim-353*, *RG-65* and *MC-30* (Panchenko et al., 2016). The two latter were excluded from our comparison according to the reasons described above. So, we used the following datasets:

- The revised version of *SimLex-999*, dubbed *RuSimLex-965* (21 word pairs remained) (Kutuzov and Kunilovskaya, 2017) and translated versions
- The translated versions of *WS353-Sim* (5 pairs remained) and *WS353-Rel* (7 pairs remained) (Panchenko et al., 2016)

4. Experimental Setup

To this point, we obtained two datasets with vectors (gaze vectors and word vectors) and one dataset with scalar values (human judgments on word similarity). For each word pair in each vector dataset, we computed cosine distance between the vectors corresponding to the words in a pair. Then for every word pair in each dataset we had a float in $\{0, 1\}$ reporting similarity between two words in this pair. In the end, to find correlation of the datasets we computed Spearman correlation value for distances between word pairs. The results of that comparison are presented in the following section. The code to reproduce the experiments as well links to the datasets and models are available at out GitHub³.

5. Results and discussion

First of all, we evaluated the gaze vector space by analyzing the nearest neighbors to certain target words. Table 1 reports the closes neighbors for Russian gaze vectors and Russian word embeddings for randomly used words (closest vectors were found with a *KD-tree* neighbor search (Maneewongvatana and Mount, 1999)). Notably, according to the notion of word relatedness, some words produced by

¹<http://rusvectors.org/en/models/>

²<http://vectors.nlpl.eu/repository/>

³<https://github.com/bakarov/subconscious-embeddings/tree/master/eye-tracking/lincr2018>

		Similarity			Relatedness		
		SimVerb	SimLex	WS353-Sim	MEN	MTurk	WS353-Rel
English	Provo , <i>gaze</i>	0.01	-0.09	-0.2	-0.09	0.19	0.07
English	GECO , <i>gaze</i>	0.06	-0.28	-0.8	-0.19	0.14	-0.6
English	Google News , <i>embeddings</i>	0.36[†]	0.35 [†]	0.77[†]	0.77[†]	0.67 [†]	0.64[†]
English	BNC , <i>embeddings</i>	0.18 [†]	0.25 [†]	-	0.76 [†]	0.69[†]	-
Russian	RSC , <i>gaze</i>	-	0.24	-0.1	-	-	0.05
Russian	Araneum , <i>embeddings</i>	-	0.39[†]	0.57 [†]	-	-	0.61 [†]
Russian	Ruscorpora , <i>embeddings</i>	-	0.28 [†]	0.74 [†]	-	-	0.57 [†]

Table 2: Performance of English word embeddings and gaze vectors across word similarity and word relatedness tasks in Spearman’s correlation value. Daggers report $pval < 0.01$, absence of a symbol report $pval > 0.05$.

		English		Russian	
		BNC	GN	Ara	RC
English	GECO	0.99[†]	0.14	-	-
English	Provo	0.97	0.23[†]	-	-
Russian	RSC	-	-	0.65[†]	0.63[†]

Table 3: Correlation of English gaze vectors with English word embeddings and Russian gaze vectors with Russian word embeddings (values report Spearman’s correlation).

gaze vectors seem to be very related to target words. So, we cannot say that the gaze vectors work well, but we also not conclude that they are just random.

Table 2 reports the correlation values for Russian and English gaze vectors (as well as word vectors) with word similarity assessments. In general, the experiments show the lack of correlation even between similarity judgments and gaze vectors, giving in most cases low correlation score. Apart from the issue of embedding evaluation, this raises the problem of whether semantic similarity is a factor affecting gaze variables during reading at all. However, due to the low size of remained word pairs in pre-processed datasets, the statistical significance of obtained results could be question, so we are not able propose any confident conclusions.

The results of pairwise comparison of distances between English gaze vectors for both eye movement datasets and embeddings vectors for both corpora are presented in Table 3 (all $pval < 0.01$). The results report very high correlation with a *BNC* model and low correlation with Google News despite the fact that a *Google News* model showed better results on word similarity task. On the other hand, variation in correlation scores for two English distributional models is high, while variation for Russian model is low. This fact possibly proves our hypothesis that eye movement data behaves differently for different languages. It is also interesting that the *Araneum* model reports the highest correlation with gaze vectors, and it also has the best results among Russian embeddings on most of word similarity tasks, while English model show an inverse pattern.

To this end, we do not say that eye movement data is worth being used as a gold standard for evaluation since we are not actually able to interpret obtained results. It is possible that substandard embeddings (that fail on word similarity task) correlates well with eye movement data (so this data is substandard), but it is also possible that word similarity data could be substandard itself, so eye movement data detects an actually good model.

6. Conclusions

In this paper, we compared word vectors, human judgments of word similarity and eye movement data of Russian and English languages. We noted that eye movement data is a some way correlated with word embeddings and even with word meaning, and the behavior of this data is different in other languages. Despite the results could be called negative, we can conclude that such data needs a more detailed investigation: that may be it would be more appropriate to use in another way of evaluation on eye movement data.

For example, one could try to train a regression model that gets word embeddings vectors and tries to predict one of the features of eye movement data. The set of words in eye movement data should be split into a train set and a test set, and an evaluation measure on test (for example, mean squared error) would report performance of word embeddings (the best embeddings should have the lowest error). The most reliable features could be selected by measuring p-value on predictions.

So, in future we plan to make such type of evaluation, trying to only adopt one of the features instead of their vector similarity. We also want to make a more extensive comparison obtaining other eye movement datasets for other languages (like the *Potsdam Corpus* (Stede, 2004)), and we plan to link other *neuro-cognitive resources* (like *fMRI* data) to word embeddings spaces, integrating current work in a big project about evaluation of word embeddings on different types of linguistic data.

Acknowledgements

We would like to thank three anonymous reviewers for their valuable comments and effort to improve the manuscript. We also thank my colleague, Andrey Kutuzov, for productive discussions on this paper.

7. Bibliographical References

- Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., and Soroa, A. (2009). A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27. Association for Computational Linguistics.
- Auguste, J., Rey, A., and Favre, B. (2017). Evaluation of word embeddings against cognitive processes: primed reaction times in lexical decision and naming tasks. In *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*, pages 21–26.
- Baker, S., Reichart, R., and Korhonen, A. (2014). An unsupervised model for instance level subcategorization acquisition. In *EMNLP*, pages 278–289.
- Batchkarov, M., Kober, T., Reffin, J., Weeds, J., and Weir, D. (2016). A critique of word similarity as a method for evaluating distributional semantic models. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 7–12.
- Bruni, E., Tran, N.-K., and Baroni, M. (2014). Multimodal distributional semantics. *J. Artif. Intell. Res.(JAIR)*, 49(2014):1–47.
- Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.
- Fares, M., Kutuzov, A., Oepen, S., and Veldal, E. (2017). Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In *Proceedings of the 21st Nordic Conference on Computational Linguistics, NoDaLiDa, 22-24 May 2017, Gothenburg, Sweden*, number 131, pages 271–276. Linköping University Electronic Press.
- Gerz, D., Vulić, I., Hill, F., Reichart, R., and Korhonen, A. (2016). Simverb-3500: A large-scale evaluation set of verb similarity. *arXiv preprint arXiv:1608.00869*.
- Halawi, G., Dror, G., Gabilovich, E., and Koren, Y. (2012). Large-scale learning of word relatedness with constraints. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1406–1414. ACM.
- Hill, F., Reichart, R., and Korhonen, A. (2016). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*.
- Hohenstein, S. (2013). *Eye movements and processing of semantic information in the parafovea during reading*. Ph.D. thesis, Universitätsbibliothek der Universität Potsdam.
- Kennedy, A., Hill, R., and Pynte, J. (2003). The dundee corpus. In *Proceedings of the 12th European conference on eye movement*.
- Kutuzov, A. and Kunilovskaya, M. (2017). Size vs. structure in training corpora for word embedding models: Araneum rassicum maximum and russian national corpus. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 47–58. Springer.
- Kutuzov, A. and Kuzmenko, E. (2016). Webvectors: A toolkit for building web interfaces for vector semantic models. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 155–161. Springer.
- Maneewongvatana, S. and Mount, D. M. (1999). Its okay to be skinny, if your friends are fat. In *Center for Geometric Computing 4th Annual Workshop on Computational Geometry*, volume 2, pages 1–8.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.
- Miller, G. A. and Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1):1–28.
- Panchenko, A., Ustalov, D., Arefyev, N., Paperno, D., Konstantinova, N., Loukachevitch, N., and Biemann, C. (2016). Human and machine judgements for russian semantic relatedness. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 221–235. Springer.
- Radinsky, K., Agichtein, E., Gabilovich, E., and Markovitch, S. (2011). A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th international conference on World wide web*, pages 337–346. ACM.
- Rogers, A., Drozd, A., and Li, B. (2017). The (too many) problems of analogical reasoning with word vectors. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017)*, pages 135–148.
- Rubenstein, H. and Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.
- Schnabel, T., Labutov, I., Mimno, D. M., and Joachims, T. (2015). Evaluation methods for unsupervised word embeddings. In *EMNLP*, pages 298–307.
- Smith, N. J. and Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.
- Søgaard, A. (2016). Evaluating word embeddings with fmri and eye-tracking. *ACL 2016*, page 116.
- Stede, M. (2004). The potsdam commentary corpus. In *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, pages 96–102. Association for Computational Linguistics.
- Wehbe, L., Vaswani, A., Knight, K., and Mitchell, T. (2014). Aligning context-based statistical models of language with brain activity during reading. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 233–243.
- Yang, D. and Powers, D. M. (2006). Verb similarity on the taxonomy of wordnet. In *The Third International Word-Net Conference: GWC 2006*. Masaryk University.

8. Language Resource References

- Cop, Uschi and Dirix, Nicolas and Drieghe, Denis and Duyck, Wouter. (2017). *Presenting GECCO: An eyetrack-*

ing corpus of monolingual and bilingual sentence reading. Springer.

Laurinavichyute, AK and Sekerina, Irina Alekseevna and Alexeeva, SV and Bagdasaryan, KA. (2017). *Russian sentence corpus: Benchmark measures of eye movements in Reading in cyrillic.*

Luke, Steven G and Christianson, Kiel. (2017). *The Provo Corpus: A large eye-tracking corpus with predictability norms.* Springer.