ISSN 2007-9737

Testing of Statistical Significance of Semantic Changes Detected by Diachronic Word Embedding

Vladimir V. Bochkarev[1,*], Yulia S. Maslennikova[1], Anna V. Shevlyakova[2]

[1] Kazan Federal University, Institute of Physics, Kazan, Russia

[2] Kazan Federal University, Institute of Philology and Intercultural Communication, Kazan, Russia

yuliamsl@gmail.com, {Vladimir.Bochkarev, AVShevlyakova}@kpfu.ru

Abstract. In recent years, methods based on word embedding models have been widely used for solving problems of semantic change estimation. The models are trained on text corpora of various years. Semantic change is detected by analyzing changes in distance between words using vector space alignment or by analyzing changes in a set of words that are most similar in meaning to a target word. Testing for statistical significance of the detected effects has not been detailly discussed in previous studies. This paper focuses on the problem of testing statistical significance of semantic change. Besides, we consider the problem of finding a confidence interval of estimates of semantic distance between words. We allow for the influence of two random factors. The first one is associated with the use of random initial conditions and stochastic optimization when training the model, the second one results from a random selection of texts for a training corpus. The proposed approach is based on the use of resampling of a training set of texts. The proposed method is tested on the COHA corpus.

Keywords. Semantic change, word embeddings, bootstrapping, corpus of historical American English.

1 Introduction

Natural languages are not fixed, they constantly evolve to reflect changes in life, culture, and human experience. Development of corpus linguistics and tools of computational analysis offered new opportunities for investigation of language evolution and change. Various studies on lexical semantic change have aimed at developing a reliable and simple automatic method for detecting changes in word meaning [1-7].

Currently, embedding-based analysis is the most popular method for studying semantic change. It is based on the hypothesis that a word distribution can be used to estimate a word meaning as the change in meaning correlates with changes in the context of use of a word [8-10]. When utilizing word embedding models, training is performed using stochastic algorithms. Thus, different training runs can provide different word embedding models. The second factor that influences the training result is a random selection of training texts.

Most previous studies have not paid sufficient attention to testing the statistical significance of the revealed changes in distribution of words. To conduct stability experiments, the embedding algorithm was run twice in [11], each time with different random seeds. Testing of semantic change significance and estimation of p-values is considered in [3].

However, it was already performed at the stage of analyzing the time series of changes in distances to identify the change point by using the Mean Shift Model [12]. Time series that were measured as the cosine-similarity between pairs of words were built in [4]; then, a hypothesis about significant change of the mean value of the obtained time series was tested.

A number of previously obtained results on change in word semantics were questioned in [13]. The authors propose to always validate discovered effects by repeating calculations on a specially created corpus containing stationary (without any changes in statistical properties over time) data.

This gave us an idea of using text set resampling (see about resampling methods in [14, 15]) to test the statistical significance of semantic changes.

Bootstrapping was applied in [16] to analyze the influence of text selection for a training corpus on variations in the cosine distance estimates obtained using word embedding, as well as variations in word neighborhoods. This work did not consider diachronic changes, experiments were conducted employing several small synchronic corpora.

After performing the analysis for a group of words, the authors conclude the work by putting forward the recommendations to never "rely on single embedding models for distance calculations, but rather average over multiple bootstrap samples, especially for small corpora".

The problem of calculating confidence intervals of semantic distances, as well as testing the significance of semantic changes, is solved for the case of using explicit word vectors in [17]. The approach proposed in it uses a bootstrap-like procedure.

In our paper, we develop a similar approach for semantic change detection methods that use word embeddings. The influence of two random factors is considered: 1) fluctuations at each new run associated with the use of random initial conditions and stochastic learning algorithms; 2) fluctuations resulting from a random selection of texts for the training corpus.

In contrast to [16], we show how to consider the degree of influence of each of the two described factors. The proposed approach is based on the resampling or bootstrapping techniques that are used to create random subsamples of the text corpus for training word embedding models.

The approach proposed in [13] allows one only to reveal or exclude some effect in large groups of words. However, our method allows one to draw conclusions about the statistical significance of semantic changes of individual selected words, as well as to find confidence intervals for estimating semantic distances between the target pairs of words and solve other similar problems.

The article has the following structure. Section 2 provides an overview of the most popular methods of diachronic analysis of word embedding models. Section 3 describes the computational method and the dataset on which it is tested. Section 4 discusses the factors that affect the range of fluctuations in estimates of semantic distances. It also describes the procedure for testing the significance of changes in the value of semantic distances. Section 5 discusses repeatability of word neighborhoods and describes a scheme for testing the significance of possible semantic changes. Section 6 discusses some English words which distribution changed significantly in 1990-2010.

2 Related Works

In recent years, analysis using word embedding models has become a widely used technique for detecting changes in meanings of words [18]. The previous works have considered two main approaches to applying word embedding to the study of semantic change. The first of these approaches assumes that word embedding models are trained on text corpora of different years, then one space is projected onto another one using the vector space alignment algorithm.

In this case, the distances between word forms (as a rule, the metric of the cosine distance) in the aligned space are used as an assessment of the change in word meanings [3]. Compared to the sequential training procedure proposed in [2], the space alignment technique (after training on corpora of different years) provides more efficient training which can be parallelized for large corpora.

This methodology is applied in [4] in a large-scale cross-linguistic analysis using 6 corpora spanning 200 years and 4 languages (English, German, French and Chinese). However, it is shown in [11] that the alignment-based approach is unstable with respect to different random seeds in the embedding algorithm. Therefore, it is less reliable when solving the problem of detecting changes in word usage.

Moreover, it is also sensitive to proper names and requires their filtering. Despite this fact, the word embedding alignment approach [4] is still prevailing and has been often used in recent studies, for example, in [19, 20]. In earlier studies, changes in word semantics were considered through changes of their relations with similar words.

For example, statistically ranked lists of verbal predicate-nominal object constructions and differences at the level of word types are examined in [1].

A graph-based approach relying on dependency parsing of sentences is proposed in [21]: the so-called distributional thesauri-based networks from The Google Book syntactic n-grams corpus are calculated for different periods of time and grouped to obtain word-centric sense clusters corresponding to different time periods.

The same can be done using word embedding models. If we have a certain method of estimating semantic similarity, we can use it to determine the groups of words that most similar in meaning to the studied one at each target time interval.

Then, the decision on the presence of semantic changes can be made based on the analysis of changes in the lists (as described in the above-mentioned works).

For example, Gonen et al. [11] train word embedding models on corpora relating to different time intervals and then compare the neighborhoods of words most similar in meaning.

If the nearest neighbours of the word are different in the two corpora, it means that the word has changed its meaning since the word embeddings reflect distribution of words. The authors note that large sets of neighbours are more stable (k = 1000) than smaller ones, which were considered in a similar work [22].

Thus, decisions about semantic changes are made either based on the analysis of changes in vectors representing words (and, accordingly, distances between such vectors) or based on the analysis of the neighborhoods of words in the semantic space.

At that, in the second case, the word neighbourhood is found using estimation of distance between the given word and other words. Therefore, first, we will consider how to check statistical significance of the detected changes in semantic distance between words.

Then, we will consider how to check statistical significance of changes in constituents of the neighborhood of a word in the semantic space.

3 Data and Method

The Corpus of Historical American English (COHA) developed at Brigham Young University [23] was used to test the proposed approach. COHA is based on the texts of approximately 107,000 documents published between 1810-2010 and contains over 400 million words. An important advantage of COHA is that it is balanced as it contains approximately the same number of documents of each genre in each decade.

The genres (for example, fiction, magazines, non-fiction, news) were selected according to the Library of Congress's categorization scheme. Part-of-speech tags are also available. Besides POS tags, COHA also contains pre-extracted word lemmas that we used to lemmatize the set of training texts. According to the creators of COHA, the corpus has an accuracy of 99.85%, which means that on average there is one error in about every 670 words.

As COHA contains carefully selected texts and is characterized by a stable number of documents of different genres over the decades, many rare words are absent from COHA, which makes it useful only for the analysis of relatively common terms. However, currently, it is one of the most requested tools for studying the evolution of the English language over the past two centuries.

Texts in COHA are aggregated by decades. Since our goal is only to demonstrate the application of the proposed method, we selected the last two decades for the analysis. Thus, we considered the changes in the distribution of words occurring between 1990s and 2000s.

COHA contains 23.5 thousand of texts relating to the interval 1990-2009. We create M random subsamples, each time selecting half of the texts from the target time interval. Thus, the size of each sub-sample approximately equals the size of the corpus in each of the compared time intervals (1990-1999 and 2000-2009).

The average size of the obtained subsamples is about 27.5 million tokens. Since the texts are of different length, the size of the subsamples is also slightly different, but the standard deviation of the length of the subsample is small and amounts to 1.22%. Since all text subsamples were obtained by random selection from a large number of texts, synchronic semantic shifts are not likely to occur.

The model was trained on each of the subsamples. The resulting set of models will further be called sample *A*. To distinguish the influence of the above-described two random factors, one more control sample is needed.

We make a random subsample of texts so that its size is as close as possible to the average size of the obtained subsamples. For this sub-sample, we also repeat the model training procedure *M* times with different random initial conditions.

The resulting set of models will be referred to as sample *B*. The difference in the results for this set of models is associated only with the randomness of the initial conditions and with the use of stochastic optimization algorithms. As for the first set of models, besides this factor, the differences also result from differences in the sets of texts used for training.

We start by selecting a pair of target words between which we want to estimate the semantic distance. Then, we calculate the cosine distances between vectors representing this pair of words for each of the models in sample *A*. Let σ_A denote the standard deviation for the obtained *M* estimates. Similarly, for models from sample *B*, we calculate *M* distance estimates for the same pair of words and find their standard deviation, which is denoted by σ_B.

The variance of distance estimates for models from sample *A* includes two terms related to the above-mentioned two random factors. The first one associated with the use of different initial conditions at each run of training, as well as with the use of stochastic optimization will be denoted by σ_{rl}^2.

The second one related to a random choice of texts from sample *A* in the process of training the models will be denoted by σ_{st}^2. Comparing the values σ_A and σ_B, we can determine each of the terms separately. Since all models from sample *B* are trained on the same set of texts, the variance in sample *B* is determined only by the first of the two factors:

$$\sigma_{rl}^2 = \sigma_B^2. \quad (1)$$

Using this formula, we find an expression for the definition and σ_{st}:

$$\sigma_{st} = \sqrt{\sigma_A^2 - \sigma_B^2}. \quad (2)$$

The Gensim library [24] was utilized to train word embedding models. The parameters of the model were chosen after a series of experiments. We used a continuous bag-of-words (CBOW) [18] model, a vector representation of words that equals 100, and a context window with a width of 5. The training was performed using function negative sampling.

Note that the minimum frequency threshold was 5, and 10^{-4} was used as the threshold for subsampling of a frequent word. The rest of the parameters were set by default. See [25, 26] for details on parameter selection techniques.

4 Analysis of Fluctuations and Changes in Semantic Distances

In our work, the value *M*=100 was chosen for the sample size *A* and *B*. This number might be excessive for practical tasks; however, our goal is to study the influence of various factors on the range of fluctuations in estimates of semantic distances.

First, we consider how the estimates of the cosine distance between words change at repeated training runs of the model. A priori, it can be assumed that the model gives the most adequate distance estimates for pairs of words with similar meaning.

We select all words that have a total frequency of at least 10 uses within the period 1990-2009, and that are used at least 5 times in each of the 100 training text sets of sample *A*. The obtained list includes 35,536 words. The mean value of the cosine distance and its standard deviation were determined for each pair of words from this list based on the control sample *B*.

Approximately 631 million different pairs of words can be formed using the above-mentioned number of words. For visual presentation, all pairs were divided into groups according to the value of the average distance with a step of 0.05 (i.e., from 0 to 0.05, from 0.05 to 0.1, etc.).

Figure 1-A shows the distribution of the standard deviation values for each of the classes as a box-and-whisker diagram.

As can be seen from the figure, the standard deviation of the cosine distance grows rapidly with an increase in the mean value of the distance; and

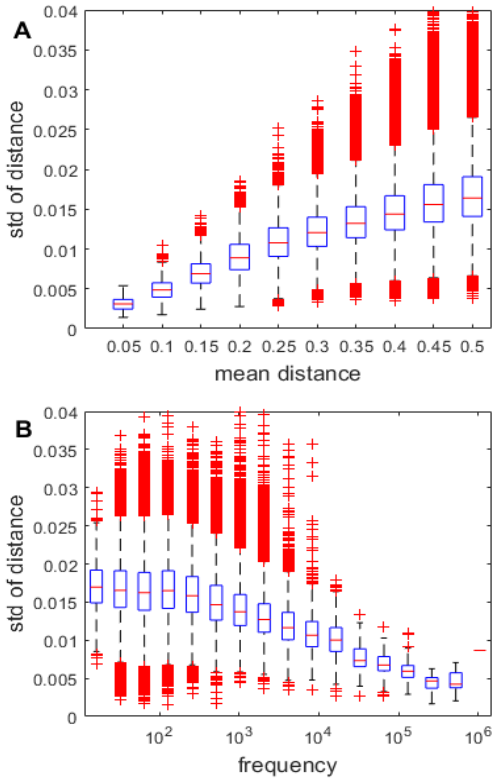

Fig. 1. A is the standard deviation of the cosine distance estimates for repeated training runs σ_{rl} in dependence to different mean cosine distance; **B** is the standard deviation of the cosine distance estimates for repeated training runs σ_{rl} in dependence to word frequency

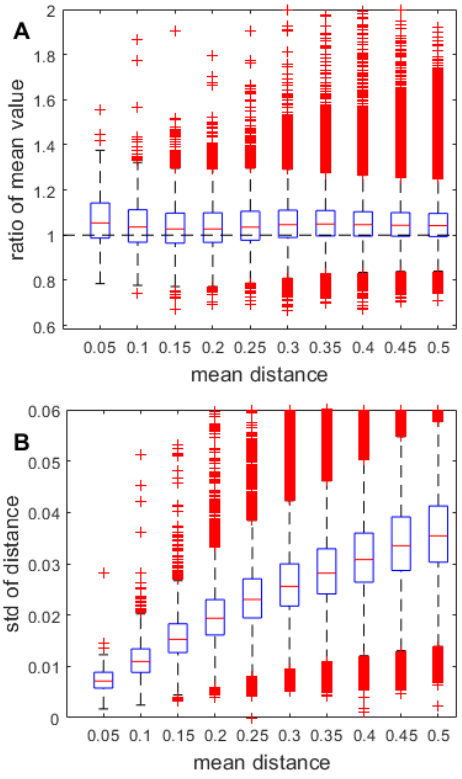

Fig. 2. A is increase in the average value of the estimate of the cosine distance when varying the set of texts in the training sample; **B** is the standard deviation of the cosine distance estimates when varying the set of texts in the training sample σ_{st} in dependence to the mean value of the cosine distance

this dependence is close to linear up to the values of approximately 0.25. On the other hand, it is seen from the figure that the range of the standard deviation for individual words is large for each of the distance ranges.

Spearman's correlation coefficient between the mean of the cosine distance and the standard deviation of its estimate is 0.287.

It is also natural to assume that the standard deviation of the distance between words depends on the frequency of those words. The vectors corresponding to the most frequent words should be better reproduced at repeated training runs; accordingly, the estimates of the distances between frequent words should have a lower standard deviation. We divide words into frequency classes by rounding down the binary logarithm of their frequency. Thus, frequencies of two words that fall into the same class differ by less than 2 times. Figure 1-B shows dependence of the standard deviation of the cosine distance estimates at repeated training runs for pairs of words with approximately equal frequency (that is, belonging to the same frequency class).

The standard deviation of the cosine distance estimates is indeed significantly dependent on frequency. However, this dependence is significantly slower than in the classic $1/\sqrt{N}$ sampling rule. For example, the median standard deviation differs only by a factor of 2.95 for classes that differ in frequency by a factor of 256 (210 and 218).

ISSN 2007-9737

Let us now consider how strongly the estimates of semantic distances vary when the composition of texts in the training set changes. To do this, we calculated the distances for each pair of words in accordance with the models from sample *A* and sample *B*. It is noteworthy that the distance estimates based on sample *A* are on average slightly higher than those based on sample *B*.

Figure 2, A shows the dependence of the ratio of the average distance using sample *A* to the average distance using sample *B* depending on the average semantic distance between words (for sample *B*). As one can see, the effect is slightly stronger for word pairs with similar semantics.

Then, the standard deviation of the estimates of the cosine distance associated with the variation of the set of texts in the training sample was determined using formulas (1, 2). Since we have only estimates of standard deviations σ_A and σ_B, and not their true values, the value inside the radical symbol in expression (2) may turn out to be negative.

However, in practice, this happens extremely rarely, only in 1,300 cases among 631 million word pairs. In such cases, we took σ_{st} equal to zero. Figure 2, B shows the dependence of the standard deviation of the estimates of the cosine distance when varying the set of texts in the training sample as a function of the mean value of the cosine distance. Like in Figure 1, A, we see a rapid increase in the standard deviation with the increase of the average distance.

Spearman's correlation coefficient between the mean of the cosine distance and the standard deviation of its estimate is 0.288. The correlation coefficient between σ_{st} and σ_{rl} is significantly higher and equals 0.6313. It means that if the distance between a pair of words fluctuates strongly from one training run to another, then it will most likely fluctuate strongly when the set of texts in the training corpus varies.

Figure 3 shows the dependence of the standard deviation on the frequencies of both words. As you can see, the standard deviation increases rapidly with the decreasing frequency of both words. To make the analysis more convenient, the same way as described above, we select the pairs of words which frequencies are approximately equal. Then we divide the set of all selected pairs into frequency classes the same way as it was done to construct Figure 1, B. The distributions of the standard deviation values for different frequency classes are shown in a box-and-whisker diagram in Figure 4, A.

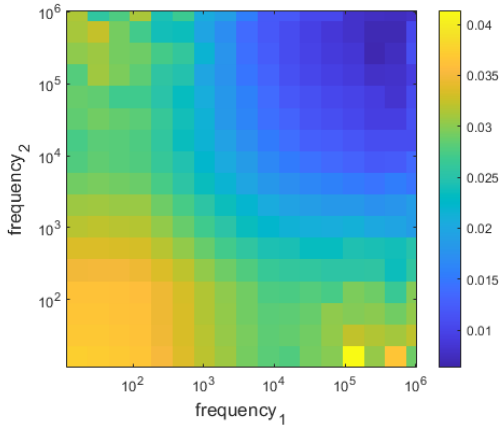

Fig. 3. Standard deviation of the cosine distance estimates when varying the set of texts in the training sample σ_{st} depend on frequency of words

If we, as described above, take the classes corresponding to the frequency values 210 and 218, the median value of the standard deviation for these classes will differ by a factor of 3.97. That is, the frequency dependence for σ_{st} is somewhat stronger than for σ_{rl}. Comparing Figures 1, B and 4, A, it is noteworthy that σ_{st} in the entire frequency range is much higher than σ_{rl}.

It is also of interest to consider the dependence of the standard deviation of cosine distance estimates on the document word frequencies. If a word occurs in a small number of documents, the frequency of its use in different contexts will fluctuate more when we use a subsample of texts.

This leads to an increase in fluctuations in the estimates of the distances between the target word and the rest ones. The analysis is carried out in the same way as described above: we select a set of pairs of words with approximately the same document frequency and divide it into classes by frequency. The results are shown in Figure 4, B.

Thus, the fact that the standard deviation of the cosine distance estimates depends on frequency (both usual and documentary) is beyond doubt, but the growth of the standard deviation with frequency decrease is significantly slower than one would expect. Now we describe how to use resampling to

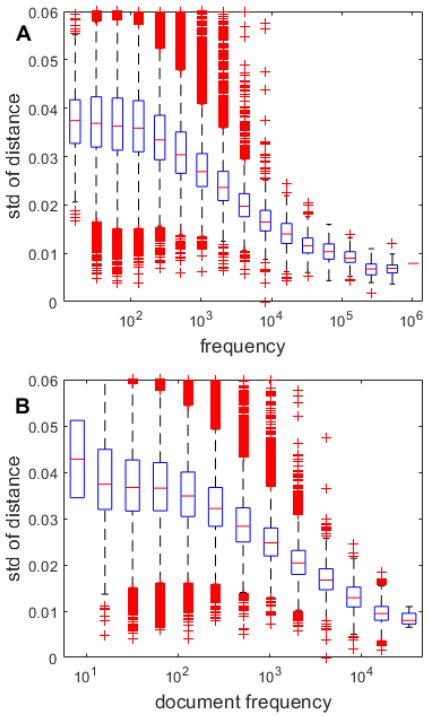

Fig. 4. A is the standard deviation of the estimates of the cosine distance when varying the set of texts in the training sample σ_{st} in dependence to frequency of words; **B** is the standard deviation of the cosine distance estimates when varying the set of texts in the training set σ_{st} in dependence to a document frequency of words

check the significance of changes in semantic distances. Suppose we have two time intervals, and the distance between the selected pair of words in the second interval changes compared to the first interval.

As described above, we create a combined corpus of texts related to both intervals and make *M* random subsamples of texts so that the size of each of them is equal to the size of the corpus for the second of the two periods.

Having trained the model on each of the *M* subsamples and calculated the distances for the target pair of words, we obtain an imitation of the empirical distribution of the distance for the case of the validity of the null hypothesis of the absence of the changes. If *M* is large enough, then one can simply calculate the percentage of cases where the distance deviation is greater than or equal to the given distance value. This calculation provides an estimate of the p-value. If the computational power is limited and *M* is not large, one can restrict oneself to calculating the standard deviation of the distance estimates and use it to check the significance of the changes.

It can be noted that the described scheme is not completely balanced: to detect changes, it is necessary to train the model 2 times (for the two compared time intervals); and to check the significance of the models, they are trained *M* times.

Following the recommendations described in [16], for each of the compared intervals, we can also repeat the training *N* times (where N is comparable to M) and average the distance estimates obtained for each of the *N* models.

This technique is similar to what is done using meta-embeddings [27]. Averaging allows one to suppress fluctuations associated with the randomness of the initial conditions and the use of stochastic optimization. Obviously, the averaged distance estimates have the standard deviation:

$$\sqrt{\sigma_{st}^2 + \frac{1}{N}\sigma_{rl}^2}. \tag{3}$$

Expression (3) can be used to check the significance of averaged distance changes. On the other hand, having two sets of *M* distance estimates on the sample *A* and *N* distance estimates for the second time interval, any suitable nonparametric criterion (for example, the Wilcoxon rank-sum test) can be used to test the significance of their differences.

Consider the frequently encountered problem of testing the significance of the difference in distance estimates in two pairs of words calculated using the same corpus. One should make subsamples of texts using the corpus of the same size as the corpus itself.

This is possible if we allow re-selection of the text as is done in classic bootstrapping. Having trained the model using each of the *M* subsamples and calculating the distance based on it, we can determine the confidence interval for the distance estimate either using the obtained empirical distribution or by calculating the standard deviation of the distance estimates and using the well-known asymptotic formulas.

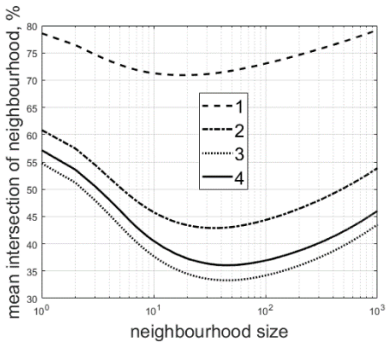

Fig. 5. Average percentage of intersection of word neighborhoods. 1 - for pairs of words from Sample *B*, 2 - for pairs of words from Sample *A*, 3 - for pairs of models where the first one was trained on the data from the interval 1990-1999, and the second one – on the data from the interval 2000-2009, 4 - for the distance estimates averaged over 100 training runs for 1990-1999 and 2000-2009

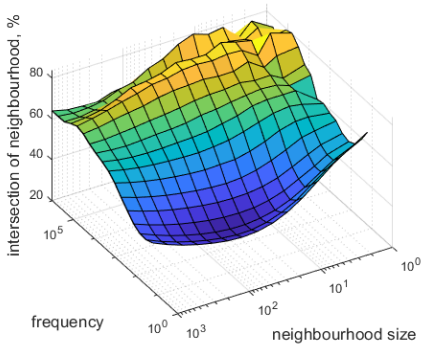

Fig. 6. Average percentage of intersection of word neighborhoods for Sample *A* in dependence to word frequency

5 Stability of Word Neighborhood

As mentioned above, many papers on detection of semantic changes compare word neighbourhoods, i.e., a list of words that are most similar in meaning to the target word. Besides changes in semantics, these lists can also change due to various random factors, including the randomness of the initial conditions and stochastic optimization in the process of training of the model, as well as due to the random selection of texts in the training corpus.

To analyse the reproducibility of the neighbourhoods of words, we used not only samples *A* and *B* but also conducted N=100 cycles of the model training for each of the periods 1990-1999 and 2000-2009. The calculation results are shown in Figure 5.

Curve 1 shows the average intersection percentage of neighborhoods found for pairs of models from sample *B*. Here we see the degree of repeatability of neighborhoods we can expect from one training run to another on the same training corpus. Curve 2 shows the average intersection percentage of neighborhood for pairs of models from sample *A*.

As one can see, taking into account a random sample of texts in the training set leads to a sharp drop in the percentage of neighborhood intersection. It should be emphasized that curves 1 and 2 refer to the case when there are no semantic changes. For example, a neighborhood of size 1 is just the word which is most similar in meaning to the target word. When we compare neighborhoods of a particular target word found using the two models, these words can either match (100% intersection) or not (0% intersection).

However, on average, the detected nearest words for the pairs of models from sample *B* intersect (as we see in Figure 5) in 78.6% of cases if a number of target words is large. As for sample *A*, the average intersection percentage drops sharply and is only 60.9%.

We compare the model trained on the texts of the first interval (1990-1999) with the model trained on the texts of the second interval (2000-2009) (Curve 3 in Figure 5).

The difference between curves 2 and 3 shows in what percentage of cases changes in the neighborhood of words result from changes in the language, not from the above-mentioned random factors. We can also determine the neighborhoods of words in 1990-1999 using distance estimates obtained by averaging over *N*=100 models for this time interval. Similar calculations will be performed using 100 models for the interval 2000-2009.

The average intersection percentage of the neighborhoods of words found in this way is shown by curve 4. As one can see, in this case, the degree of intersection of the neighborhoods is slightly higher, which creates better conditions for detecting actual semantic changes.

Figure 6 shows the average percentage of intersection of the word neighborhoods for sample

A, depending on the word frequency. As one might expect, the intersection percentage significantly depends on frequency (more rare words show decrease in neighbourhood repeatability). Thus, we can quantitatively describe the effects mentioned in [13].

Let us describe a scheme for testing the statistical significance of changes in the composition of word neighborhoods. Suppose we have two time intervals, and the neighborhood of the target word changes in the second interval compared to the first one.

As described above, we create a combined corpus of texts related to both intervals and make *M* random subsamples of texts so that the size of each of them is equal to the size of the corpus for the second of the two periods. We determine the neighborhoods of the target word for the models trained on these subsamples.

Assume we are interested in a neighborhood of size *n*, and the neighborhoods for the two time intervals contain *m* common words. We will choose different pairs from *M* samples and determine the intersection of the neighborhoods of the target word for them. The number of common words for the *i*-th pair is denoted by m_i. The null hypothesis is that, in fact, there are no changes in distribution, and the observed differences result from random factors. Under this assumption, the p-value is simply found as the percentage of pairs that satisfy the condition:

$$m_i \leq m. \tag{4}$$

To conclude this section, we consider for what percentage of words the neighborhood changes are statistically significant within the interval 1990-2009. Using the method described above, we checked statistical significance of changes in the neighbourhoods (of the size from 1 to 1000) of each word from each of the 100 pairs of models trained for the intervals 1990-1999 and 2000-2009. Figure 7 shows average percentage of words for which we revealed statistically significant changes (according to the level of significance of 0.05) in the neighborhood of the semantic space.

The figure shows an interesting phenomenon: large neighborhoods turn out to be unstable and change significantly for the majority (89.8%) of words. Thus, it is hardly advisable to use them to analyze changes in word meanings. This

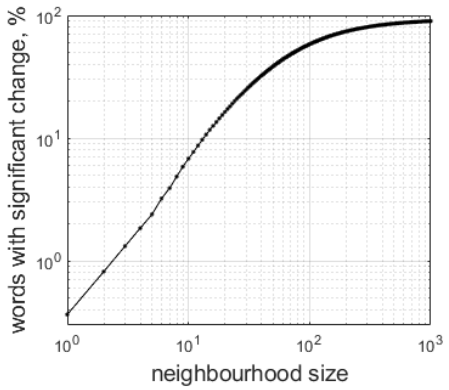

Fig. 7. Average percentage of words that show significant changes of neighbourhood in the semantic space in dependence to a neighborhood size (for the time interval 1990-2009)

phenomenon requires further study. Significant changes are recorded for 0.37% of words for the neighbourhoods of size 1 and for 0.82% for the neighbourhood of size 2, respectively.

Note that these facts cannot be used to make conclusions about a number of words whose meanings really changed in the considered time interval. The number of such words can exceed the mentioned one, however, the used data do not allow drawing a conclusion about changes in their meanings.

6 Case Study

We selected a group of words that showed significant changes (the level of significance is 0.05) in distribution both for neighbourhood 1 and neighbourhood 2 when comparing the intervals 1990-1999 and 2000-2009. There were ten words included in the list.

We did not consider two of the words as one of them probably results from tokenization errors and another one is a rare polysemantic abbreviation. Changes in the distribution of the rest 8 words is discussed in detail in this section. The only frequent word from this list is the word windows. According to the Google Books Ngram data, its relative frequency in 1990-2009 was $9.63 \cdot 10^{-5}$.

The other words are relatively rare. Their relative frequencies in the same time peri-od varied

Table 1. Nearest neighbours of the selected words

Nearest neighbours in 1990-1999	Nearest neighbours in 2000-2009
bora	
Dora, Canberra, mara, Isadora, angora, serra, Madera, Italia, pandora, calamari	Tora, angora, Omar, Dahmer, Yemen, Dora, kali, guerrilla, Afghanistan, Basra
Headnote	
storybook, guides, guide, fabled, recipes, Seuss, sidebar, guidebook, illustrated, guided	sidebar, trends, outdoorsy, hiking, redbook, invigorating, outdoors, creates, lifestyle, ultimate
katrina	
Kat, Katie, Kathie, mom, Kathy, Katy, Sabrina, Rosie, Irina, Katya	hurricane, fema, evacuee, Rita, hurricanes, disaster, Orleans, Louisiana, devastate, devastation
lassitude	
solitude, plenitude, vicissitude, fortitude, barrenness, moodiness, callousness, vastness, restlessness, exhilaration	vicissitude, bandar, helplessness, restlessness, weightlessness, hopelessness, insolence, eagerness, timelessness, lucidity
playoff	
midseason, threegame, preseason, postseason, singleseason, season, regularseason, nfc, game, firstround	postseason, season, regularseason, preseason, midseason, singleseason, sevengame, game, tigers, firstround
spf	
sunscreen, uv, mm, cm, iu, moisturizer, deg, pf, ml, sunblock	moisturizer, moisturize, moisturizing, lotion, gel, serum, sunscreen, pf, mascara, gloss
Windows	
PCs, desktop, browser, Macintosh, software, PC, Microsoft, CPU, CDROM, Pentium	XP, linux, PC, desktop, Symantec, PCs, windowpane, firewall, window, software
XP	
thoughtless, unrepentant, thus, remorseless, artless, selfless, grievous, lifeordeath, torturous, incapable	Windows, pc, pcs, linux, Symantec, software, Microsoft, MSN, firewall, desktop

between $4.92 \cdot 10^{-7}$ and $4.72 \cdot 10^{-6}$. The words are shown in Table 1. The two columns show 10 nearest neighbours of each of the selected words for 1990-1999 and 2000-2009. Though we consider the neighbourhoods of size 1-2, we decided to show a larger neighbourhood (see Table 1) of the considered words to make it easier to estimate the degree of change in their distribution.

The first word is *bora*. It is a polysemantic word that have several meanings. For example, it is 1) a kind of cold wind, 2) a sacred site for rituals held by native peoples of Australia and 3) a part of names (such as geographical or personal, etc.). *Bora* is associated with the nearest neighbours in the following way. In 1990s, *bora* is strongly associated with the word *Dora*. Dora Bora is a resort island (as well as a hat name created on this island) Pandora (fish) and calamari are a seafood that can be prepared on the island.

Associations with *mara* and Canberra are due to the second meaning of the word, a sacred site for Australian native peoples. Bora is a north to north-eastern katabatic wind in areas near the Adriatic Sea, so it combines with the word *Italy*, a country in this geographical area. Serra Bora, Isadora Bora are names. Madera Bora is a firm producing wooden items.

The distribution changed in 2000s. *Bora* started associating with the word *Tora*; Tora Bora is a cave complex in Afghanistan that is known for a stronghold location of the Taliban. Therefore, associations with guerilla and Basra can be explained by the wars related to these places. Yemen is often compared to Tora Bora and Yemen Bora is also a name. The other name associated with Bora is Omar, probably Omar Mohammad, one of the Taliban leaders.

The word *bora* was used in the 2000s twice as often than in 1990s. Significant changes in distribution of this word are obvious in 2000s; however, no new meanings are revealed, one of the previously emerged meanings is foregrounded.

One more example is the word *headnote*. It means "a brief summary, comment, or explanation that precedes a chapter, report, etc.". Therefore, in 1990s *headnote* associates with storybook, guide, guides, recipe, Seuss (American writer and illustrator Dr. Seuss, the full name is Theodor Seuss Geisel), guide-book, illustrated, guided.

All these words relate to print production and the associations are clear. In 2000s the distribution changed. It still associated with printed production (redbook (the book and a bookshop), ultimate (law)), the description and creation of headnotes (outdoorsy, trends, creates, life-style, invigorating). *Headnote* also associates with outdoors and hiking.

However, the revealed distribution is strongly associated with COHA errors when the machine does not see punctuation marks. Development of computer industry explains that *headnote* is associated with sidebar, it is still a piece of additional information placed on the screen but not in pages of a book or a magazine. The word *headnote* became ten times more frequent in 2000s than in the previous decade. Its distribution changed though no new meanings emerged.

The distribution of the word *Katrina* is also an interesting example. Originally, it is a woman name. In 1990s, this word is associated with other woman names like Kat, Sabrina, Rosie. In 2000s, the distribution changed due to Hurricane Katrina. Most of the neighbours relate to the *hurricane*, the places where it occurred and the disaster that it brought. The frequency of use of Katrina in 2000s increased compared to the previous decade. The

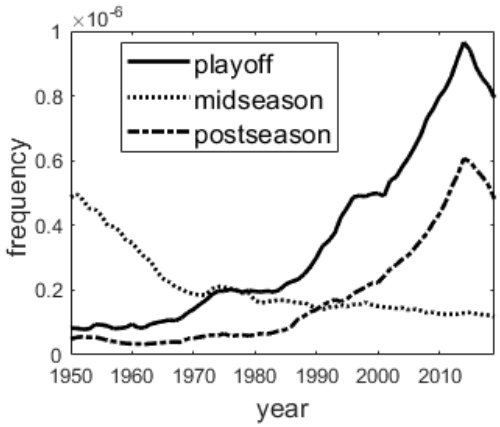

Fig. 8. Frequency of the words *playoff, midseason* and *post-season* in the American English subcorpus of Google Books Ngram in 1950-2019

denotata of the word changed, therefore we can say that a new meaning emerged.

If we compare the two decades, the word *lassitude* also changed its distribution. It means a "condition of weariness or debility characterized by lack of interest, energy, or spirit". Though there are variations in the distribution throughout the decades, it still means the same negative state of mind and body. Changes in the associations might result from cultural changes.

The word *playoff* that relates to sport events also shows insignificant variations in distribution and does not change its meaning in this context. However, we observe changes in the order of the nearest words. They are probably due to changes in frequency. For example, frequency of the words *midseason* and *postseason* changes which causes changes in the associations.

Figure 8 shows frequencies of the words *play off*, *midseason*, *postseason* obtained using the Google Books Ngram corpus. The word *Windows* was already widely used as the name of an operating system in the 1990s.

If we look at Table 1, we can see that the first and second nearest words in the list in that period are PCs and desktop. The rest words in the list also refer to the computer field. In 2001, Microsoft released a new version of the operating system called *Windows XP*, which was extremely successful. We see that in the 2000s, the word most strongly associated with *Windows* is *XP*.

The second word in the list is the name of the competing operating system *Linux.* Thus, this example shows that the word *Windows* has not gained a new meaning, and the changes in distribution are associated with the release of a successful version of the well-known operating system *Windows*. Thus, the distribution of this word changes with time, however, the meaning relates to computer production and use.

It seemed interesting to analyse the distribution of abbreviations. Often, they denote different things and relate to various aspects of life and science.

One of the obtained abbreviations is *XP*. In the 2000s, *XP* was used in most cases in the corpus as a part of the *Windows XP* compound name (in fewer cases - as a part of the *Athlon XP* compound name); therefore, *Windows* appears to be the first word in the list of the nearest ones. As one can see from the table, all words in the list strongly associated with *XP* in the 2000s relate to computer software application.

It should be noted that the word form *XP* was previously used in English, sometimes as a part of compound names. Often, *XP* has been formed as an abbreviation of certain phrases, including such words as experimental, experience, extreme, extended, etc. However, before 2000, this word form is found in only a few sources of COHA.

In the 1990s there are only a few uses of this word in a single book, as an abbreviation for the name of a genetic disease. We see that the distribution of *XP* has changed; and in the last target decade it was mostly used as relating to the computer application sphere.

Among the selected words is the abbreviation *SPF* that very often implies sun protection factor though have some other meanings as spray polyurethane foam etc. In 1990s, it is associated with the words *UV, sunscreen, sunblock, moisturizer* that may refer to the SPF-factor of sunscreen lotions.

IU, mm, cm, ml, and *deg* can refer both to sunscreens and other things abbreviated by *SPF*. In 2000s, the distribution shows that the word is mostly used in sphere of cosmetology meaning *SPF-factor* (moisturizer, lotion, gel, mascara). Therefore, one of the previously existed meaning has become dominant.

To conclude, we note that manual check showed that it is true that 7 of the selected 8 words show significant change in distribution. However, only two words can be regarded as ones that have obtained new meanings. Thus, the considered cases confirm one more time that change in word distribution does not always cause change in word semantics.

7 Conclusion

This paper proposes an approach that allows one to test significance of semantic changes that are detected using word embeddings. The key idea is to use resampling of a set of texts in a training sample. The proposed method allows testing of statistical significance of changes in estimates of semantic distances, as well as of changes in a list of the nearest neighbours of target words. It also can be used to find boundaries of a confidence interval for estimates of semantic distances.

Using the resampling and bootstrapping techniques indeed requires a multiple increase in the amount of computation. However, currently, the required computing power is no longer the main limiting factor for solving problems of semantic change detection.

Rather, further progress in this study area is limited by the lack of effectiveness of the existing models. In any case, we faced no difficulties in performing the required calculations using an ordinary personal computer.

We also used resampling to investigate what factors determine the range of variation in semantic distance estimates. The influence of two random factors was considered: the use of random initial conditions and stochastic optimization when training the model, and a random selection of texts for the training corpus.

Unlike previous works, for example [16], we indicate a way to consider the degree of influence of each of these factors separately. Though there is increase in the standard deviation accompanied by decrease in word frequency, it is much slower than could be expected a priori.

We also analysed how the above-mentioned two random factors affect variations in the composition of the list of the nearest neighbors of words and estimated the stability of neighborhoods

of different sizes. It was shown that the use of large neighborhoods to detect semantic changes is impractical.

Finally, we analysed a number of words for which the fact of changing the nearest neighbors in the semantic space seems to be the most reliable and reasonable from the statistical point of view. We selected 8 words for which, in accordance with the calculations carried out in our work, the change of the nearest two neighbours during the transition from the 1990s to the 2000s can be considered statistically significant.

The manual analysis showed that significant change in the distribution of these words at this time is undeniable except for one word. However, only one word gained a new meaning. The case study confirmed one more time that changes in distribution does not always mean change in a word meaning.

Acknowledgments

This work has been funded by Russian Science Foundation, grant № 20-18-00206.

References

1. **Cavallin, K. (2012).** Automatic extraction of potential examples of semantic change using lexical sets. Proceedings of the 11th Conference on Natural Language Processing, pp. 370–377.
2. **Kim, Y., Chiu, Y. I., Hanaki, K., Hegde, D., Petrov, S. (2014).** Temporal analysis of language through neural language models. Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science, pp. 61–65. DOI: 10.3115/v1/W14-2517.
3. **Kulkarni, V., Al-Rfou, R., Perozzi, B., Skiena, S. (2015).** Statistically significant detection of linguistic change. Proceedings of the 24th International Conference on World Wide Web, pp. 625–635. DOI: 10.1145/2736277.2741627.
4. **Hamilton, W. L., Leskovec, J., Jurafsky, D. (2016).** Diachronic word embeddings reveal statistical laws of semantic change. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Vol. 1, pp. 1489–1501. DOI: 10.18653/v1/P16-1141.
5. **Kutuzov, A., Øvrelid, L., Szymanski, T., Velldal, E. (2018).** Diachronic word embeddings and semantic shifts: a survey. Proceedings of the 27th International Conference on Computational Linguistics, pp. 1384–1397.
6. **Rodina, J., Trofimova, Y., Kutuzov, A., Artemova, E. (2021).** ELMo and BERT in semantic change detection for Russian. Lecture Notes in Computer Science, Vol. 12602, pp. 175–186. DOI: 10.1007/978-3-030-72610-2_13.
7. **Pivovarova, L., Kutuzov, A. (2021).** RuShiftEval: a shared task on semantic shift detection for Russian. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference Dialogue 2021, Vol. 20, pp. 1–21.
8. **Harris, Z. (1970).** Papers in structural and transformational linguistics. Dordrecht, Reidel. DOI: 10.1007/978-94-017-6059-1.
9. **Rubenstein, H., Goodenough, J. B. (1965).** Contextual correlates of synonymy. Communications of the ACM, Vol. 8, No. 10, pp. 627–633. DOI: 10.1145/365628.36565.
10. **Firth, J. R. (1957).** A synopsis of linguistic theory, studies in linguistic analysis 1930-1955. Special volume of the Philological Society, pp. 1–32.
11. **Gonen, H., Jawahar, G., Seddah, D., Goldberg, Y. (2020).** Simple, interpretable and stable method for detecting words with usage change across corpora. 58th Annual Meeting of the Association for Computational Linguistics, pp. 538–555. DOI: 10.18653/v1/2020.acl-main.51.
12. **Taylor, W. A. (2000).** Change-point analysis: A powerful new tool for detecting changes. Taylor Enterprises.
13. **Dubossarsky, H., Weinshall, D., Grossman, E. (2017).** Outta control: Laws of semantic change and inherent biases in word representation models. Proceedings of the 2017 Conference on Empirical Methods in

Natural Language Processing, pp. 136–1145. DOI: 10.18653/v1/D17-1118.

14. **Efron, B. (1981).** Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. Biometrika, Vol. 68, pp. 589–599. DOI: 10.1093/biomet/68.3.589.
15. **Good, P. (2006).** Resampling methods. A practical guide to data analysis, 3rd Ed. Birkhäuser Basel.
16. **Antoniak, M., Mimno, D. (2018).** Evaluating the stability of embedding-based word similarities. Transactions of the Association for Computational Linguistics, Vol. 6, pp. 107–119. DOI: 10.1162/tacl_a_00008.
17. **Bochkarev, V., Shevlyakova, A. (2021).** Calculation of a confidence interval of semantic distance estimates obtained using a large diachronic corpus. Journal of Physics: Conference Series, Vol. 1730, p. 012031. DOI: 10.1088/1742-6596/1730/1/012031.
18. **Mikolov, T., Chen, K., Corrado, G. S., Dean, J. (2013).** Efficient estimation of word representations in vector space. International Conference on Learning Representations.
19. **Yao, Z., Sun, Y., Ding, W., Rao, N., Xiong, H. (2018).** Dynamic word embeddings for evolving semantic discovery. Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, pp. 673–681. DOI: 10.1145/3159652.3159703.
20. **Schlechtweg, D., Hatty, A., Del-Tredici, M., Walde, S. S. I. (2019).** A wind of change: detecting and evaluating lexical semantic change across times and domains. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 732–746. DOI: 10.18653/v1/P19-1072.
21. **Mitra, S., Mitra, R., Riedl, M., Biemann, C., Mukherjee, A., Goyal, P. (2014).** That's sick dude!: Automatic identification of word sense change across different timescales. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pp. 1020–1029. DOI: 10.3115/v1/P14-1096.
22. **Wendlandt, L., Kummerfeld, J. K., Mihalcea, R. (2018).** Factors influencing the surprising instability of word embeddings. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 2092–2102. DOI: 10.18653/v1/N18-1190.
23. **Davies, M. (2012).** Expanding horizons in historical linguistics with the 400-million word corpus of historical American English. Corpora, Vol. 7, No. 2, pp. 121–157. DOI: 10.3366/ cor.2012.0024.
24. **Řehůřek, R., Sojka, P. (2010).** Software framework for topic modelling with large corpora. Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pp. 45–50. DOI: 10.13140/2.1.2393.1847.
25. **Levy, O., Goldberg, Y., Dagan, I. (2015).** Improving distributional similarity with lessons learned from word embeddings. Transactions of the association for computational linguistics, Vol. 3, pp. 211-225. DOI: 10.1162/tacl_a_00134.
26. **Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., Dean, J., (2013).** Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems, pp. 3111–3119. DOI: 10.48550/arXiv.1310.4546.
27. **Yin, W., Schütze, H. (2016).** Learning word meta-embeddings. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp. 1351–1360. DOI: 10.18653/v1/P16-1128.

Article received on 12/09/2024; accepted on 21/10/2024.
**Corresponding author is Vladimir V. Bochkarev.*