Calculation of a confidence interval of semantic distance estimates obtained using a large diachronic corpus

Vladimir V. Bochkarev, Anna V. Shevlyakova
Kazan Federal University, Kazan, Russia
vbochkarev@mail.ru, anna_ling@mail.ru

Abstract. Several methods for detection changes in words semantics and appearance of new word meanings have been suggested. These methods use different techniques of estimating semantic distance between words. They are based both on neural network vector models and on simpler vector representations that use frequencies of n-grams including the studied words. This paper proposes a method for calculation the confidence interval of the semantic distance estimations obtained based on the frequency data of n-grams extracted from the large diachronic corpus. This task is complicated because the question about the law of distribution of frequency fluctuations of words and n-grams, despite a number of studies, remains open. The confidence intervals are calculated by statistic modeling using random permutations of n-gram frequencies. To test the proposed method, estimation of semantic distance between two Russian synonyms is used as an example.

1. Introduction
Vector representations of word semantics have been widely used in different fields of science. They allow one to estimate semantic distance between words by using large sets of linguistic data. These methods are based on the distributional hypothesis. The general idea of this hypothesis is that there is a correlation between distributional similarity and meaning similarity [1,2,3]. Therefore, word distribution can be used to estimate its meaning. There are different algorithms of distributional meaning acquisition [4,5,6]. Currently, the most widely used methods are based on neural network vector models. However, simpler representations based on n-gram frequencies are also used to solve various tasks such as studies of language evolution, detection of new meanings of words and other problems of computational linguistics.

At the same time, the issues concerning statistical justification of data obtained by this method are not sufficiently highlighted in the available literature. This article describes a method for automatic calculation of the confidence interval of semantic distance estimates based on n-gram frequency analysis.

To test the method, the Google Books Ngram corpus is used. It includes data on frequencies of words and word combinations in eight languages and covers the time interval of five centuries [7]. The Russian subcorpus of Google Books Ngram contains data of word and n-gram frequencies based on the texts of more than one million books published between 1486 and 2019 (the third version of the corpus was used). The total amount of words in the subcorpus is 89 billion that several times exceeds the number of words in any other Russian corpus.
2. Method
At the first stage, all contexts of the compared words are to be found. In this case, the distribution of the word can be numerically characterised by frequency vectors of its usage in these contexts. The Google Books Ngram corpus contains frequency data on 1-, 2-, 3-, 4- and 5-grams. Frequency data on 5-grams are used in some works to train neural network vector models to solve various problems such as studying changes in word semantics [8,9]. There is some difficulty as the corpus does not contain n-grams which total frequency is less than 40 within the entire period of its use. Therefore, the corpus may not include 5-, 4- and even 3-grams that contain rare words. For this reason, data on frequencies of 2-grams (a smaller context region consisting of only a couple of consecutive words) are used in our paper. First, let us found all word combinations of the \( W_x \) and \( xW \) forms in the corpus (where \( W \) is the studied word and \( x \) is some word) and denote the set of words ‘\( x' \) (that are found in these word combinations) by \( I(W) \). To estimate the difference in the distribution of the two words \( W_1 \) and \( W_2 \), the co-occurrence frequency vector of these words is found (for all the combinations with the words ‘\( x' \) of the set \( I(W_1) \cup I(W_2) \)). It is natural to normalize the resulting vectors to 1 since they represent the probability distribution of the word usage in the 2-gram context.

To estimate differences in the vectors that correspond to different words (or one and the same word in different time intervals), it is required to choose a certain metrics. The Jensen-Shannon divergence (JSD), that is widely applied in computational linguistics [10,11], will be used as an example in this paper. However, the suggested method can be applied to any metrics from the list of those used for analysing differences in probability distributions.

Using the described vector representation made on the basis of a sufficiently large diachronic corpus, one can analyze how the difference in the distribution of words changes over time. Let us consider the example. Figure 1 shows the change in the semantic distance between the Russian words aktivnee and aktivnei (more active). These words are synonyms and represent the comparative form of the adjective aktivnii (active). There is a slight stylistic difference in the use of this pair of words. The suffix -ee (in aktivnee) makes the word sound bookish and the suffix -ei (in aktivnei) is more associated with colloquial speech.

![Figure 1. The change in the distance (JSD) between the vectors that characterise the Russian words aktivnee and aktivnei](image)

It is seen that the semantic distance estimate varies greatly over time. This can be due to the following reasons. The distribution is converging as the language becomes more "democratic" and the form with the colloquial suffix -ei has been increasingly used in book speech reducing stylistic differences. On the other hand, before making any meaningful conclusions, it is necessary to check whether the observed changes in the distance are statistically significant, and whether they are not associated with any factors that are not related to the language evolution, for example, with a change in the annual corpus size. This
relates to significant difficulties since the question of the probabilistic distribution of sample frequencies of words remains open. This question is discussed, for example, in [12,13,14].

When the law of frequency distribution is unknown, it is possible to use bootstrap-like procedures to simulate empirical frequency distributions. Suppose one can choose a time interval during which the semantics of the word $W$ does not undergo significant changes, and the corpus size also does not change significantly. For example, the interval 1960 – 2010 can be chosen for the example shown in Figure 1. Under the assumption made above, changes in the frequencies of the word combinations $Wx$ and $xW^r$ within the selected interval are due to random fluctuations. Bootstrapping can be used to simulate empirical distribution for each of these word combinations [15]. The suggested procedure includes the following steps:

- The interval within which the frequency distribution of word combinations (including the studied word) does not presumably change is selected.
- Since the corpus size can vary within this interval, relative frequencies of the word combinations are calculated. To do this, empirical frequencies of the word combinations in a certain year are normalized by the total size of the corpus in that year.
- For each component of the frequency vector (that is, the frequency of a certain word combination), one of the annual values of the relative frequency from the chosen time interval is selected independently of the other components.

Generating random vectors for two compared words in the described way, the Jensen-Shannon divergence between them is calculated. This bootstrap-like procedure allows one to find a confidence interval for estimating the semantic distance between words.

3. Result

Figure 2 shows distributions of the semantic distance between the words aktivnee and activnei, calculated based on the data from two time intervals: 1960 – 1991 (solid line) and 1928 – 1953 (dash-dotted line).

![Figure 2](image)

Figure 2. The distance distribution (JSD) between the vectors that characterise the Russian words aktivnee and activnei (based on the data from 1928 – 1953 and 1960 – 1991 time intervals)

For the 1960 – 1991 interval, the average value of the Jensen-Shannon divergence is 0.337 and the standard deviation is 0.0301. Having calculated the mean and standard deviation of the annual values of the semantic distance for the same time interval based on the empirical data (see Figure 1), the values 0.317 and 0.0318 are obtained, respectively. This good agreement between the model and empirical values indicates that the distribution of the words under consideration has slightly changed over the years. Let us find the 95% confidence interval for estimating the semantic distance. Its boundaries are equal to 0.284 and 0.405. For the 1928 – 1953 interval, the obtained average value of the Jensen-Shannon divergence is 0.621 and the standard deviation is 0.0983. The boundaries of the 95% confidence are 0.445 and 0.829.
For the 1928 – 1953 interval, the mean value is 0.956 and standard deviation is 0.193 based on the empirical data. The agreement between the model and empirical values is good for the mean value; however, the standard deviation obtained using the empirical data is 1.42 times higher than that obtained by model calculations. It is natural to assume that this results from changes in the distribution of the studied pair of words in this period. The ratio of the empirical value of the standard deviation to its model value can apparently be used to detect the nonstationarity of the frequency dynamics. This, however, requires additional experiments.

Let us consider the influence of the corpus size on estimation of the semantic distance. If the corpus size decreases, the relative value of fluctuations of the sampled frequencies increases. Aside from that, many rare word combinations that include the word under study may not be included in the corpus in this case. This leads to the increase in the estimates of the semantic distances (see the discussion of this question in [6]) and increase in their dispersion. Let us show how to take this effect into account within the framework of the proposed approach. Suppose one has data on the frequencies of word combinations (that include a studied word) for a certain time interval. It is needed to generate a vector representing the word for a different time interval, when the absolute frequency of the word is less than in the first case (for example, due to the smaller size of the corpus during this period). Following the method described above, a vector representing the word is generated, composed of the relative frequencies of the word combinations that include the given word. Multiplying this vector by the average value of the absolute frequency of the word during this period, the expected values of the frequencies of the word combinations are obtained and must further be rounded down.

As for the considered example, the average annual size of the corpus in 1960 – 1991 is 1.19 billion words but it includes 0.203 billion words in 1928 – 1953 that is 5.87 times less. Let us consider the hypothesis that the distribution of the words aktivnee and aktivnei in both considered time intervals is the same in reality, and large values of the semantic distance for the first time interval are associated with the insufficient size of the corpus in these years. Using the 1960 – 991-year data, the average values of the semantic distance with a reduced size of the corpus by one proportion or another is calculated. Figure 3 shows that the average value of the distance estimate increases with a decrease of the corpus size (the ratio of the corpus size in the model to the real size of the corpus in 1960 – 1991 is plotted horizontally).

![Figure 3](image)

Figure 3. The average value of the estimated distance (JSD) between the vectors characterizing the words aktivnee and aktivnei at the corpus size decrease

The calculation results show that if the corpus size is 0.203, the expected average distance value is 0.384. It is slightly higher than the distance value in 1960 – 1991. However, it is not enough to explain extremely high distance values in 1928 – 1953.

Thus, when shifting from the 1928 – 1953 interval to the 1960 – 1991 interval, a statistically significant decrease in the semantic distance is observed between the Russian words aktivnee and
aktivnei that cannot be explained by random fluctuations or a change in the corpus size. This leads to the conclusion that the distribution of these words converged in the 20th century.

4. Conclusion
Despite the development of neural network vector models of word semantics, methods based on frequencies of n-grams are still successfully used to solve various tasks. This paper proposes a method that allow one to determine boundaries of the confidence interval of semantic distance estimates obtained based on n-gram frequencies extracted from the large diachronic corpus. Since the question of the probability distribution of fluctuations of words and word combinations remains open, the empirical distribution of semantic distance estimate is simulated using a bootstrap-like procedure.

To show the possibilities of the proposed method, the evolution of the distribution of two Russian synonyms (aktivnee and aktivnei) is considered as an example. This example shows how to find the boundaries of the confidence interval of the semantic distance estimation, as well as how to calculate the p-value for a statistical hypothesis about the difference in the semantic distance between these words for two time intervals.

The proposed method can be used to study language evolution. For example, it can be applied in algorithms for detecting new meanings of words, as well as for a wide range of computational linguistics problems in which it is required to evaluate semantic similarity of words.

Acknowledgements
This research was financially supported by Russian Science Foundation, grant № 20-18-00206.

References
[1] Harris Z 1970 Papers in structural and transformational Linguistics (Dordrecht: Reidel) p 775
[2] Rubenstein H and Goodenough J 1965 Communications of the ACM 8 (10) 627
[3] Pantel P 2005 Inducing ontological co-occurrence vectors. Proc. of the 43rd Conference of the Association for Computational Linguistics (Ann Arbor, June 2005) (Morristown, NJ, USA: ACL) pp 125-132
[4] Sahlgren M 2008 Italian Journal of Linguistics 20 (1) 33
[5] Bullinaria J A and Levy J P 2007 Behavior Research Methods 39 510
[6] Wartena C 2013 CEUR Workshop Proceedings 986 8
[7] Michel J-B, Shen Y K, Aiden A P, Veres A, Gray M K, et al. 2011 Science 331(6014) 176
[8] Mitra S, Mitra R, Riedl M, Biemann C, Mukherjee A and Goyal P 2014 That's sick dude!: Automatic identification of word sense change across different timescales Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics (Baltimore, USA, June 23-25 2014) (Long Papers vol 1) (Baltimore, USA: ACL) pp 1020-1029
[9] Kulkarni V, Al-Rfou R, Perozzi B and Skiena S 2015 Statistically Significant Detection of Linguistic Change Proc. of the 24th Int. Conf. on World Wide Web (May 18–22, 2015, Florence, Italy) (Geneva, CHE: IW3C2) pp 625-635
[10] Endres D M and Schindelin J E 2003 A new metric for probability distributions IEEE Transactions on Information Theory 49(7) 1858
[11] Dagan I, Lee L and Pereira F 1997 Similarity-Based Methods for Word Sense Disambiguation Proc. of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conf. of the European Chapter of the Association for Computational Linguistics (Madrid, Spain, July 1997) (Baltimore, USA: ACL) pp 56-63
[12] Baayen R H 2001 Word Frequency Distributions (Dordrecht, Netherlands: Springer)
[13] Baayen R H and Lieber R 1996 Comput. Hum. 30(4) 281
[14] Bochkarev V V and Shvelyakova A V 2019 J. Phys.: Conf. Ser. 1391 012053
[15] Efron B 1981 Biometrika 68(3) 589