Semantic similarity and analysis of the word frequency dynamics

To cite this article: V V Bochkarev et al 2017 J. Phys.: Conf. Ser. 936 012067

View the article online for updates and enhancements.
Semantic similarity and analysis of the word frequency dynamics

V V Bochkarev, Yu S Maslennikova, A A Svetovidov
Kazan Federal University, Kremlevskaya str.18, Kazan 420018, Russia
E-mail: svetovidov1994@gmail.com

Abstract. In this study a similarity in changes of frequencies dynamics for semantically related words was analyzed using word statistics extracted from more than 4.5 million books written over a period of 205 years. The approach is based on the correlation analysis of 1-grams frequency dynamics. We analyzed the frequencies correlation of synonym pairs, their corresponding antonymous groups and random words pairs. Also, we compared several metrics to find the most effective for assessing the degree of similarity in the dynamics of use of different words. Comparing differences between logarithmic rank variations in pairs of synonyms and random word pairs, significant differences are found, though they are smaller than it could be expected.

1. Introduction

Evolution of written language is significantly influenced by the wide range of cultural and historical factors. Social and political events also influence on stylistics and semantic context of written speech. Furthermore, choice of specific word forms in the text is determined by grammar requirements. With the lapse of time, these external factors bring certain contribution into lexicon usage statistics, as some word forms become more prevalent for use, while the others become less preferable. Therefore, it is expected that related in meaning concepts and word forms will have similar word usage frequency dynamics. However, obsolete concepts and archaisms may have opposite usage frequency trends to concepts, which replace the first ones in use. Today an increasing number of works are dedicated to statistical analysis of semantically related word frequencies. For example, the approach called “relative cosine similarity” is proposed in [1]. Distributed word vectors were analyzed, the cosine distance was used to estimate the synonymic similarity for different part-of-speech separately. This approach gave a significant improvement the accuracy of synonyms recognition.

The similar study is proposed in [2], where a synonym acquisition trajectory is analyzed, chosen by non-native English speakers learning English. It is found that with the growth of language proficiency the number of correctly selected synonyms increases, but even advanced learners face difficulties with synonyms in context-called-for unique constructions. The study [3] develops the functional systems theory (politics, religion etc.), analyzing also the frequency graphs for words «money», «power» and «love». It was noticed that frequency changes trends for concepts «love» and «money» are rather close to each other in English. However, the authors analyzed the words without consideration of their synonyms or words from the same semantic groups. Therefore, our research could supplement achieved results of the above work. Recent results [4] shows that semantically related words show strong phase coherence. It is noted that found patterns in the statistics of language may be a consequence of changes in the cultural framework that influences the thematic focus of writers.
Authors of this study analyzed only a core vocabulary of 5,630 common English nouns using only correlation metrics. The purpose of this work is to figure out, whether there is a correlation between frequency usage changes of semantically related words based on pre-verified lists, and also to determine the most informative measure frequency series proximity.

Our study is based on the 2012 version of the Google Ngram database [5]. We focus on 1-grams data, which consist of single word frequency counts for every year independently. The 2012 version of the database is annotated with parts-of-speech tags [6], which allow the extraction of particular word classes. Additionally, WordNet database was used [7]. WordNet is one of the largest lexical databases, which provides access to more than 140 000 words from 4 parts of speech (nouns, adjectives, verbs, adverbs) and the information about their semantic relationships. For our study, we considered groups of common nouns and adjectives.

2. Methods
For English language, our starting database was the 1-gram counts from 1800 year because, although the available data start at 1520, the data for the first 300 years is rather sparse. Using the parts-of-speech tag included in the database [6], we extracted the 1-gram information for the noun and adjective classes. This procedure finally left a core vocabulary of 493 nouns and 584 adjectives. Figure 1 demonstrates examples of normalized time series of the group of synonyms and antonyms.

![Figure 1](image1.png)

**Figure 1** Frequency dynamics of synonyms (on the left) and antonyms (on the right)

![Figure 2](image2.png)

**Figure 2** $L_\infty$-metric between random pairs of adjectives with different words frequencies

To acquire information about semantic similarity degree through comparison of word-use data, correlation analysis was executed. Following list of pairwise distances was used: Euclidean distance ($L_2$-metric), City block metric ($L_1$-metric), Chebychev distance ($L_\infty$-metric), cosine distance, Pearson’s correlation and Spearman’s rank correlation. Correlation metrics and cosine distance only reflect changes in forms of researched series, however, distances for the first three metrics also depend on
absolute values of frequencies. We have been interested in the analysis of words frequency dynamics, that’s why time series of frequencies was normalized by the mean for the considered period (1800-2005). Also, we compared logarithmic rank variations, i.e. logarithms of numbers from a word frequencies list, as in [4].

To estimate the significance of acquired distances between semantically related words, estimation of distances between random words (reference pairs) were also executed. Reference pairs from similar parts of speech were chosen, because several papers demonstrate steady difference in usage frequency trends, for example, between nouns and adjectives, which should be taken into consideration [6]. While choosing pairs of random words the average usage frequencies were considered. It is well known that time series of rare words frequencies are extremely noisy. As it is shown in [8], relative value of words frequency variations has a power-law dependence on frequency: $\sigma / f \sim 1 / f^{0.316}$. This fact leads to increasing of acquired distance values between rare words frequencies. Figure 2 shows, how the average Chebychev distance varies according to changes of logarithms of random adjective pairs usage frequencies. Therefore, reference pairs with words frequencies that are similar to analyzed synonym and antonym groups were selected for our research.

3. Results
At the first step, similarities between the normalized frequencies dynamics of synonyms and reference pairs were estimated. However, the analysis of the distribution for distances between groups of synonyms shows that it is almost same as the distribution of distances between random pairs of words. Differences that are more significant were obtained by comparing the logarithmic rank variations as in [4]. For each pair of synonyms (the total number of pairs was 1457 for adjectives and 1484 - for nouns), we selected 40 pairs of randomly selected words of the same part of speech with sufficiently close frequencies. Then, the logarithmic ratio of distances in reference pairs and pairs of synonyms was calculated. In Figure 3 distributions of the logarithmic ratio using $L_1$-metrics are shown for adjectives and nouns separately. The positive values of the logarithmic ratio show that distances between synonyms is less than distances in reference pairs. The observed asymmetry of both distributions (with a significant prevalence of positive values) indicates that semantically related words are close to each other with comparison of reference random pairs.

![Figure 3. Distributions of the logarithmic ratio of distances between reference pairs and pairs of synonyms for adjectives (on the left) and nouns (on the right)](image)

To estimate the significance of the found effect, the null hypothesis about the homogeneity of samples in reference pairs and pairs of synonyms was tested. According to the sign test, $p$-value was $6.84 \cdot 10^{-191}$ for adjectives and $3.48 \cdot 10^{-203}$ for nouns. Different metrics were compared based on analysis of the accuracy for semantically related words recognition using different pairwise distance metrics. Thus, it can be concluded that the found effect is significant.

To compare different metrics, we estimated the accuracy of semantically related words recognition. To do this, we compared distances in synonyms pairs $D_{syn}$ with distances in reference pairs $D_{ref}$ and considered the number of cases, when $D_{syn} < D_{ref}$. If the analyzed pair consists of semantically unrelated words, then the number of cases will be any value from the interval from 0 to $N$, where $N$ is the number of reference pairs. If analyzed pair consists of semantically related words, then this number
will be larger with high probability, as follows from Figure 3. In our study, the threshold value for make the decision was chosen so that errors of the first and second kind would be equal. The results for adjectives and nouns are shown on the Figure 4. We can see that the L_1-metric is the most accurate and shows minimal errors: 0.29 - for adjectives and 0.32 - for nouns.

![Figure 4. Errors of semantically related words recognition for different pairwise distance metrics](image)

4. Conclusion

Comparing distances between the series of normalized frequencies of semantically related words and similarly obtained distances for random word pairs, no significant differences are observed. On the contrary, comparing differences between increment series of ranks in pairs of synonyms and random word pairs, significant differences are found, though they are smaller than it could be expected. Apparently, a sort of competition between words and word substitution in groups of synonyms play an important role in dynamics of language use. It should also be taken into account that there are words which on average don’t have significant change in frequency of use over the last two centuries (for example, the word “unbending” in Figure1). When assessing the degree of similarity in the dynamics of use of different words, the most effective metrics are L_1 and L_2.

This work was supported by the Russian Foundation for Basic Research, Grant no. 15-06-07402. The research of the first author was supported by the Russian Government Program of Competitive Growth of Kazan Federal University.

References

[1] Leeuwenberg A, Vela M, Dehdari J, Genabith J A 2016 Minimally Supervised Approach for Synonym Extraction with Word Embeddings The Prague Bulletin of Mathematical Linguistics 111-142
[2] Liu D, Zhong S 2016 L2 vs. L1 use of synonymy: An empirical study of synonym use/acquisition Applied Linguistics. 37 Issue 2 239-261
[3] Roth, Steffen et al. 2016 The Fashionable Functions Reloaded: An Updated Google Ngram View of Trends in Functional Differentiation (1800-2000) In: Mesquita, A. (Ed.) Research Paradigms and Contemporary Perspectives on Human-Technology Interaction. Hershey: IGI-Global
[4] Montemurro M A, Zanette D H 2016 Coherent oscillations in word-use data from 1700 to 2008 Palgrave communications
[5] Michel J, Shen Y, Aiden A et al. 2011 Quantitative Analysis of Culture Using Millions of Digitized Books Science 331 (6014) 176-182
[6] Lin Y, Michel J-B, Aiden EL, Orwant J, Brockman W and Petrov S 2012 Syntactic annotations for the google books ngram corpus, in 'Proceedings of the ACL
[7] Fellbaum C. 1998 WordNet: An Electronic Lexical Database Cambridge, MA: MIT Press
[8] Bochkarev V, Solovyev V, Wichmann S 2014 Universals versus historical contingencies in lexical evolution J. R. Soc. Interface 11: 20140841