Generation of a dictionary of abstract/concrete words by a multilayer neural network

V D Solovyev, V V Bochkarev, S V Khristoforov
Kazan Federal University, Kazan, Russia
maki.solovyev@mail.ru

Abstract. Large dictionaries of abstract/concrete words were compiled for several languages by interviewing native speakers. The Russian dictionary contains only one thousand words. This article proposes a new method for automatic generation of a large (tens of thousands of words) Russian dictionary of abstract/concrete words by using neural networks trained on the Google Books Ngram corpus. Estimates of the quality of the dictionary compiled using this method are obtained. The correlation coefficient between the estimates obtained using a neural network for the level of word concreteness and the estimates obtained based on the native speakers’ responds is 0.778.

1. Introduction

It is important to distinguish between abstract and concrete words in various studies in linguistics, psychology, neurophysiology, and clinical medicine. A critical review of the latest theories on abstract concepts is presented in [1]. There is neither strict definition of abstract/concrete in linguistics nor formal method that can allow classifying words based on this characteristic. Therefore, there is a need for creation dictionaries of abstract/concrete lexemes. Large dictionaries which indicate the degree of word abstractness/concreteness have been created for the English language by interviewing native speakers [2,3]. The dictionary described in [3] contains 40 thousand words and is available for free on the Internet. To obtain estimates of abstractness/concreteness for each word in the process of compilation of this dictionary, at least 30 respondents were interviewed. Thus, more than 1200 responds were obtained. It is obvious that such method of compiling a dictionary is time-consuming and expensive. A corpus-based method of automatic generation of the dictionary of abstract/concrete words was introduced in [4]. Spearman’s correlation coefficient of automatically extracted and expert estimates was 0.7 [4].

The objective of this project is to compile a similar dictionary for the Russian language. Estimates were obtained for 1000 words based on the native speakers’ response. The respondents (university students) were offered to estimate the degree of abstractness/concreteness of words on a scale from one to five. No preliminary instructions were given. Not less than 40 estimates were obtained for each word and averaged. The created dictionary is available at the project site [5].

In the future, it is planned to expand it to 2000 words. However, to create a several-fold larger dictionary using expert estimates is too labour-consuming. Thus, artificial intelligence methods are used to automatically generate estimates based on the text corpora data. The new method used for automatic generation of dictionaries of abstract/concrete words and a Russian dictionary created by this method that includes 88 thousand words (available on the project website at the above address) is described in [6]. The method is based on the assumption that abstract words combine mainly with abstract words and concrete words combine mainly with concrete words [7]. This method is used in [8] for the English
language and the Spearman correlation coefficient was calculated for automatically obtained estimates and expert estimates from [3]. The coefficient equaled 0.71. Thus, they failed to significantly improve the result presented in [4]. Therefore, there is still a problem of automatic compilation of a dictionary that shows high agreement with the dictionary created based on the expert estimates.

This article proposes a different approach to this problem - the use of neural networks for automatic generation of the Russian dictionary of abstract/concrete words.

2. Data and Method

Two data sets were used for the experiments: 1) the dictionary that contains 1000 words (nouns) for which the values of word concreteness indices were obtained by expert assessment, 2) the dictionary of 88 thousand words (nouns and adjectives) for which the word concreteness indices were automatically obtained using the algorithm described in [6]. The number of words in the first dictionary is too small. Therefore, the words whose concreteness indices were estimated by the experts, as well as their inflectional forms (a total of 10300 different word forms) are used as the test and training samples. The lists of the inflectional forms were obtained using the OpenCorpora morphological dictionary [9].

Data on the word distribution are used as input data to obtain the estimate of the abstractness index. According to [10], a word is represented by a vector of frequencies of its use in word combinations (bigrams) with the most frequently used Russian words. Data on frequencies of the bigrams are extracted from the Russian sub-corpus of Google Books Ngram [11].

The Russian sub-corpus of Google Books Ngram is currently the largest Russian language corpus. It includes data on texts written between 1607-2009 and contains 67.1 billion words. The post-revolutionary period is represented most fully. We used the data for the period 1920-2009 to avoid problems associated with the spelling reform of 1918. The total number of words in that period is 62.3 billion.

A list of 20 thousand of the most frequent (in 1920-2009) words of the Russian language was selected. For each studied word \( W \), 40 thousand bigrams of the form \( W + X \) and \( X + W \) are potentially possible, where \( X \) is the word from the list of the most frequent words. The word \( W \) is represented by the frequency vector of these bigrams. As a rule, most of these bigrams are not used in the language (or are not found in the corpus) and thus have frequencies equal to zero. Nevertheless, it is advisable to use a uniform representation for all studied words by vectors of fixed length. Frequency vectors are normalized to 1. As it was shown in [10], using the word representation analogous to the one described above allows obtaining good results for the task of parts of speech recognition and identification of word grammatical characteristics.

The model is a classic feedforward neural network. The dimensions of each of the 4 layers were 64, 128, 128, and 1, respectively. The ELU activation function [12] was used for the hidden layers and the activation of the output layer is linear. During the training process, the neural network minimized the L1 norm of the difference between the estimation vector and the target vector.

The model was trained using the back-propagation method. The total number of parameters in the model was approximately 2.5 million. Their fitting was performed using the stochastic gradient optimization algorithm Adam [13]. It fits the parameter of the training speed using the previous values of the gradient norm for each weight, which allows one to train deep models for a fairly small number of iterations. In addition, the norm of the obtained gradient can be artificially limited in the training process so that, when it hits on the “steep” sections of the objective function, the weight increments remain within the acceptable limits.

Since the dimension of the input frequency vector is high, the number of weights of the 1st hidden layer is very large, which can lead to retraining of the model even at the beginning of optimization. To prevent this effect, a dropout layer [14] with parameter 0.5 is used after the input layer. Thus, only 20 thousand (on average) components of the input vector are used in the training at a time. This allows one to get an analogue of stochastic regularization in the training process.
To implement the training of deep neural networks, the Keras deep learning library was used [15]. The trained models were tested on test samples. The test samples included 20% of examples from the used data sets.

3. Results

The resulting Spearman correlation coefficient between the concreteness index and its neural network assessment was 0.856. It was obtained for the model trained on an automatically generated dictionary of 88 thousand words [6] during testing. However, it is of great interest to compare the obtained estimates with the index estimates provided by the experts. In this case, the resulting Spearman correlation coefficient is 0.695, which slightly differs from the values obtained in the earlier works [4,8].

More interesting results were obtained for the model that was trained using the dictionary compiled using expert assessments (10,328 of word forms). The Spearman coefficient between the concreteness index of a word from a dictionary and its neural network estimate was 0.778, which significantly exceeds the values obtained in [4,8] (0.7-0.71). Figure 1 shows the two-dimensional distribution density of the index values and their neural network estimates.

![Figure 1. Two-dimensional distribution density of the index values and their estimates](image)

It is natural to assume that the error range in the index estimation depends on a word frequency. A rare word scarcely occurs in the corpus and the selective frequencies of the bigrams that contain this word fluctuate greatly. This should lead to a greater level of error in the estimates of the concreteness index of rare words.

The correlation coefficient of the index and its estimates for words are calculated depending on their frequency. For a certain selected frequency, a group of words was selected whose frequencies differ no more than 1.28 times (or $e^{0.25}$) from the selected frequency in one direction or another. For a selected group of words, the Spearman correlation coefficient between the value of the concreteness index in the dictionary and its neural network assessment was estimated. The obtained results are shown in Figure 2. The horizontal axis shows the absolute frequency of words, i.e. the total number of word usage in the corpus in 1920-2009.
As can be seen from the figure, good estimates can be obtained for the words that are used in the used corpus more than $10^3$ times. The correlation coefficient is 0.855-0.874 for the most frequent words (with a frequency of $10^6$ uses and above).

It was noted in [3] that the Spearman correlation coefficient between two manually compiled English dictionaries [2] and [3] is 0.92. A similar comparison of two enquiries was performed for the sample of 100 Russian words. The Spearman correlation coefficient was 0.88. Apparently, this value can be considered the maximum attainable value in the computer estimation of the index. The obtained values of the correlation coefficient are quite close to this value.

![Figure 2](image_url)

**Figure 2.** The Spearman correlation coefficient between the value of the concreteness index in the dictionary and its neural network assessment

The dashed line shows approximation of the empirical values of the correlation coefficient by a power-law dependence of the form

$$1 - af^{-\beta}$$

Here $f$ is the absolute word frequency in the corpus. The exponent $\beta$ is equal to 0.0852, and the constant value is 0.494. Using this approximation, one can evaluate how accurately the model describes the concreteness index of words depending on the relative frequency of this word and the size of the corpus used.

4. Conclusion

The problem of automatic compilation of Russian dictionaries of abstract/concrete words was considered in the article. To find the concreteness index of a word, data on word combinability were used as input data. Data on frequency of bigrams that contain the studied words were used in the research. The Google Books Ngram corpus was used as a study material. The feedforward neural network (with 4 layers) was trained on the dictionary of abstract/concrete words compiled by interviewing native speakers.

The correlation coefficient between the neural network estimates and the expert estimates was 0.778 on the test sample (20% of the initial sample which was not used in training the network). This significantly exceeds the best results achieved previously. To obtain reliable estimates of a word concreteness index, the number of its uses in the corpus should be about 1 thousand. In the case of large text corpora, the proposed method allows compiling large high-quality dictionaries of abstract/concrete words using less effort, as well as performing diachronic studies.

Acknowledgements
This research was financially supported by the Russian Government Program of Competitive Growth of Kazan Federal University, and by RFBR, grants № 17-29-09163, 19-07-00807.

References

[1] Borghi A, Binkofski F, Castelfranchi C, Cimatti F, Scorolli C and Tummolini L 2017 The challenge of abstract concepts Psychological Bulletin 143(3) 263–292

[2] Coltheart M 1981 The MRC Psycholinguistic database Quarterly Journal of Experimental Psychology 33(4) 497–505

[3] Brysbaert M, Warriner A B and Kuperman V 2014 Concreteness ratings for 40 thousand generally known English word lemmas Behavior research methods 46(3) 904 – 911

[4] Snijellia B, Généreux M and Kuperman V 2019 Historical evolution of concrete and abstract language revisited Behavior research methods 51 1693–1705

[5] Technology of creation of semantic electronic dictionaries 2019 https://kpfu.ru/tehnologiya-sozdaniya-semanticheskih-elektronnyh.html

[6] Solovyev V D, Ivanov V V and Akhtiamov R B 2019 Dictionary of abstract and concrete words of the Russian language: a methodology for creation and application Journal of Research in Applied Linguistics 10 215-227

[7] Naumann D, Frassinelli D and Schulte im Walde S 2018 Quantitative semantic variation in the contexts of concrete and abstract words Proc. of the 7th Joint Conf. on Lexical and Computational Semantics (New Orleans, LA, USA) ed Nissim M, Berant J and Lenci A (Association for Computational Linguistics) pp 76–85

[8] Ivanov V and Solovyev V. 2020 Ranking concrete and abstract words using Google Books Ngram data Journal of intelligent & fuzzy systems Pre-press 1-9

[9] Bocharov V V, Alexeeva S V, Granovsky D V, Protopopova E V, Stepanova M E and Surikov A V 2013 Crowdsourcing morphological annotation. Computational Linguistics and Intellectual Technologies 12 (1) 109-115

[10] Khristoforov S, Bochkarev V and Shevlyakova A 2020 Recognition of parts of speech using the vector of bigram frequencies Analysis of Images, Social Networks and Texts (Kazan) ed W Aalst et al. (Cham: Springer) chapter 13 pp 132-142 doi: 10.1007/978-3-030-39575-9_13.

[11] Michel J-B, Shen Y K, Aiden A P, Veres A, Gray M K, et al. 2011 Quantitative analysis of culture using millions of digitized books Science 331 176-182

[12] Clevert D-A., Unterthiner T and Hochreiter Sepp 2015 Fast and accurate deep network learning by exponential linear units (ELUs) arXiv:1511.07289v5

[13] Kingma D P and Ba J 2014 Adam: A method for stochastic optimization arXiv:1412.6980v9

[14] Srivastava N, Hinton G E, Krizhevsky A, Sutskever I and Salakhutdinov R 2014 Dropout: a simple way to prevent neural networks from overfitting Journal of Machine Learning Research 15(1) 1929-1958

[15] Chollet F et al. 2015 Keras https://keras.io