Modelling of word usage frequency dynamics using artificial neural network

Yu.S. Maslennikova, V.V. Bochkarev, D.S. Voloskov

Kazan Federal University, Kremlevskaya str.18, Kazan 420018, Russia

E-mail: JSMaslennikova@kpfu.ru, vbochkarev@mail.ru

Abstract. In this paper the method for modelling of word usage frequency time series is proposed. An artificial feedforward neural network was used to predict word usage frequencies. The neural network was trained using the maximum likelihood criterion. The Google Books Ngram corpus was used for the analysis. This database provides a large amount of data on frequency of specific word forms for 7 languages. Statistical modelling of word usage frequency time series allows finding optimal fitting and filtering algorithm for subsequent lexicographic analysis and verification of frequency trend models.

1. Introduction

The word set of a language is a complex dynamical system in which words can be originated, evolve, and die out. Quantitative studies of natural languages have brought to significant progress in the understanding of word statistics [1] and language evolution [2]. A comparatively less explored question concerns the dynamics of word usage. Some representative examples include the study of distributions of n-grams in books written over the past 200 years [3], identifying trends in word frequency dynamics [4]. The model in [4] research is illustrated using a 108-word database from an online discussion group and a 1011- word collection of digitized books. The model revealed a strong relation between changes in word dissemination and changes in frequency.

The motivation for this research is based on the following objectives. In the paper [2] great attention is devoted to the impact analysis of various historical and cultural events on a word usage dynamics. Word frequency variations take place both due to external and internal factors. To make the impact analysis of external factors more justified, various prediction models can be used. Prediction models obtained in equable conditions allow to separate variations produced by regular internal reasons.

Trend analysis of the word usage dynamics (i.e. predictable components) allow to create smoothing methods like Kalman type filter. It is particularly important for filtering of usage frequency fluctuations of rare words.

high-quality prediction model for word usage frequencies could be used not only for direct forecasting, but also to back-forecasting. Similar estimates of word usage frequencies for previous periods permit to analyze mutual impact of different languages or their origins. It is important task for linguistic genetics (languages history).

Thanks to comprehensive Google Books digital library and The Google Books Ngram corpus (*http://books.google.com/ngrams/*), new possibilities for the quantitative analysis are available. This database doesn`t allow a morphological analysis, but frequency dictionaries of word-forms are available [5]. Statistical modelling of word usage frequency time series allows finding optimal fitting and filtering algorithm for subsequent lexicographic analysis and verification of frequency trend models.

For many years the field of time series forecasting has been largely analyzed by various statistical approach. Over the past 20 years in the field of nonlinear dynamical system an artificial neural networks (ANN) theory has gained more importance. In comparison with commonly used linear regression algorithms such as ARX, ARIMA this approach has several advantages. According to the Universal approximation theorem the standard multilayer feed-forward network with a single hidden layer, which contains finite number of hidden neurons, is a universal approximator among continuous functions on compact subsets of *Rn*, under mild assumptions on the activation function. [6].

The multilayer feedforward network can be trained for function approximation (nonlinear regression) or pattern recognition. The process of training a neural network involves tuning the values of the weights and biases of the network to optimize network performance, as defined by the network performance function. The most commonly used performance function for feedforward networks is mean square error (MSE) - the average squared error between the network outputs and the target outputs [7]. However, in cases when error distribution function differs from the normal distribution, MSE performance function is not optimal. In this paper a generalized approach for a neural network training based on the maximum-likelihood method is proposed.

2. Statistical model of word usage frequency dynamics

Frequency dictionaries only for accessory words were analyzed. Frequency dictionaries of The Google Books Ngram corpus are available for the time period of 1700-2008 year. For this research only 1900-2008 years period was used because it contains the most reliable data. Statistical model was created for infrequent accessory words such as '*during*', '*another*', '*against*', '*down*', '*still*', '*too*', '*per*', '*since*', '*never*', '*back*', '*thus*', '*take*', '*less*', '*himself*', '*again*', '*few*', '*among*', '*though*', '*last*' and other. Examples of the variation of word frequency in this dataset are shown in Fig. 1.

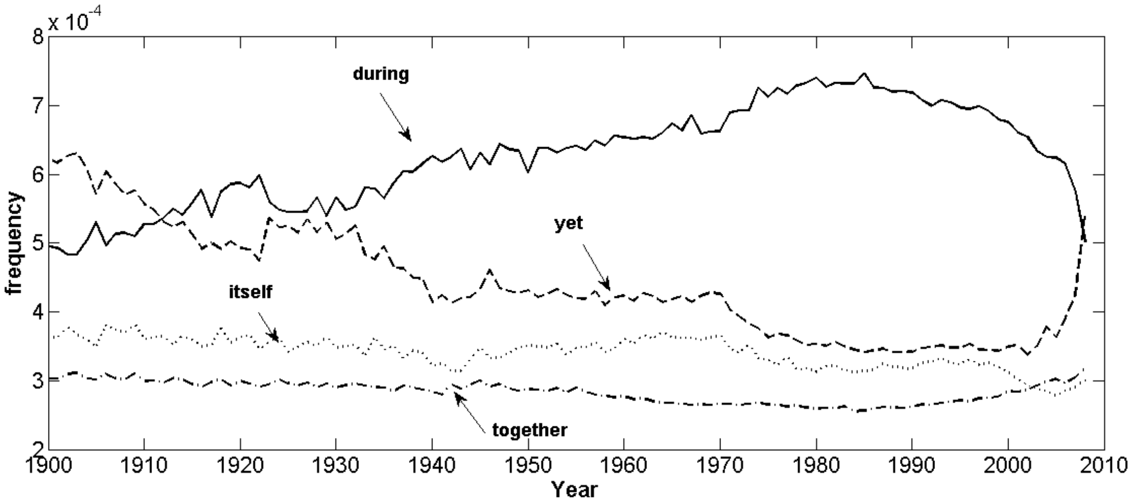

Figure 1. Frequency dynamics for example accessory words. The frequency of a word is computed as the number of occurrences of the word relative to the total number of words in a year.

As it is seen from Fig. 1, word frequency time series have various dynamics for different accessory words. That’s why for statistical modelling of these dynamics it is necessary to use a neural network with a large amount of hidden layers and neurons. Training of such complicated network could be sufficiently problematical because of the overfitting problem. In this research two-step model is proposed. As a first step, all accessory words are divided into a set of clusters based on analysis of dynamics similarity. As a second step, neural network prediction models are optimized for each cluster separately.

2.1. Word usage frequency data clustering

The approach based on AR-coefficient analysis was used for the word usage frequency data clustering. In an *AR* model of order *p*, the current output is a linear combination of the past *p* outputs plus a white noise input. The weights on the *p* past outputs minimize the mean-square prediction error of the autoregression. If *y[n]* is the current value of the output and *x[n]* is a zero-mean white noise input, the *AR(p)* model is [8]:

$$\sum_{k=0}^{p} a[k]y[n-k] = x[n].$$

These *a[k]* coefficients could be used as a input vector for the clustering. The order *p* of AR-model was chosen *p=5*. Self-organizing map neural network (SOM-net) with 2×2 neuron configuration was used for clustering of word usage frequency data. SOM-nets learn to classify input vectors according to how they are grouped in the input space. They differ from competitive layers in that neighboring neurons in the self-organizing map learn to recognize neighboring sections of the input space. Thus, self-organizing maps learn both the distribution (as do competitive layers) and topology of the input vectors they are trained on. A hexagonal grid was used to map the neurons. SOM-net wraining was performed using the Kohonen's self-training "winner-gets-all" algorithm and produced a map in which nodes spatially coincide with major accumulations of vectors in the initial attribute space and similar objects are located in neighbouring nodes [9]. As a result, all accessory words were grouped into 4 clusters. Typical trends for ascertained clusters are proposed in Fig. 2.

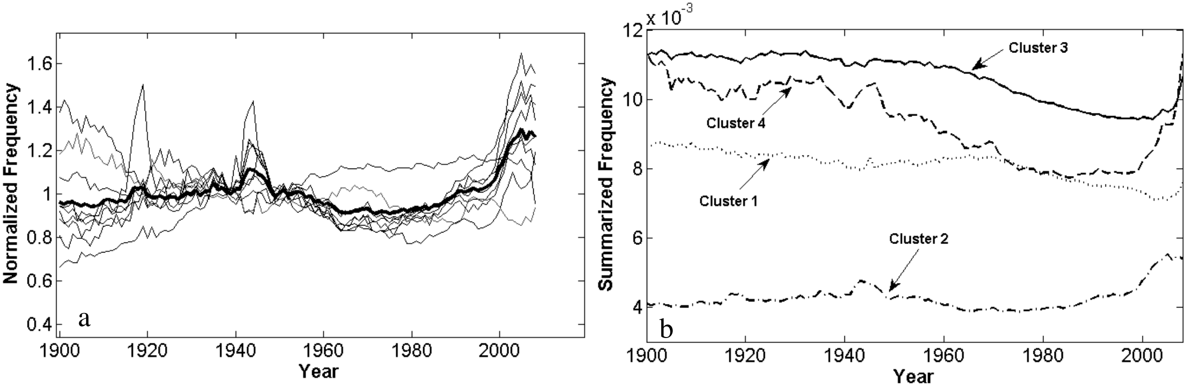

Figure 2. Frequency dynamics for clustered data: typical trend is a bold curve (a); typical trends of founded clusters (b).

2.2. Prediction model for clustered data

As it was shown early, MSE performance function for training of a neural network could be non-optimal. Generalized approach is based on the maximum-likelihood method. For infrequent accessory words error distribution function is close to the Poisson distribution. The Poisson distribution is most often invoked for rare events. As opposed to the binomial distribution the Poisson rate may actually be any number. The real contrast is that the Poisson distribution is asymmetrical: given a rate *r = 3*, the range of variation ends with zero on one side, but is unlimited on the other side. The Poisson distribution, as a data set or as the corresponding curve, is always skewed toward the right, but it is inhibited by the zero occurrence barrier on the left. The degree of skew diminishes as *r* becomes larger, and at some point the Poisson distribution becomes to the eye about as symmetrical as the normal distribution [8].

Let's define λ- parameter as a function of previous data - $\lambda_i = f(\boldsymbol{X}_i)$, then the expected value of a Poisson-distributed random variable is $Y_i \sim P(f(\boldsymbol{X}_i))$. So, the likelihood function could be defined as following:

$$l(Y,\theta) = \sum_i Y_i \log(f(\boldsymbol{X}_i)) - \sum_i f(\boldsymbol{X}_i) - \sum_i Y_i!$$

During the network training it is possible to maximize the likelihood function. For training multilayer feedforward networks, any standard numerical optimization algorithm can be used to optimize the performance function. In our case the performance function is negative likelihood function $-l(Y,\theta)$. The most commonly used backpropagation algorithm modifies the weights of a neural network in order to find a global minimum of the error function [6]. For optimizing of the performance function the Levenberg-Marquardt algorithm was used during the training of neural networks. Feedforward neural network for each clusters contained one input layer of 10 neurons and one hidden layer of 7 neurons. Examples of predicted dynamics *'against', 'down', 'during'* and *'back'* words using neural networks with MLE optimization are proposed in Fig 3 with results of linear regression neural networks (LR) and neural networks with MSE performance function. LR neural networks showed STD = $1.01 \cdot 10^{-5}$ (a standard deviation). The conventional neural network (with MSE performance function) showed STD = $0.90 \cdot 10^{-5}$, neural networks with MLE performance function showed STD = $0.46 \cdot 10^{-5}$.

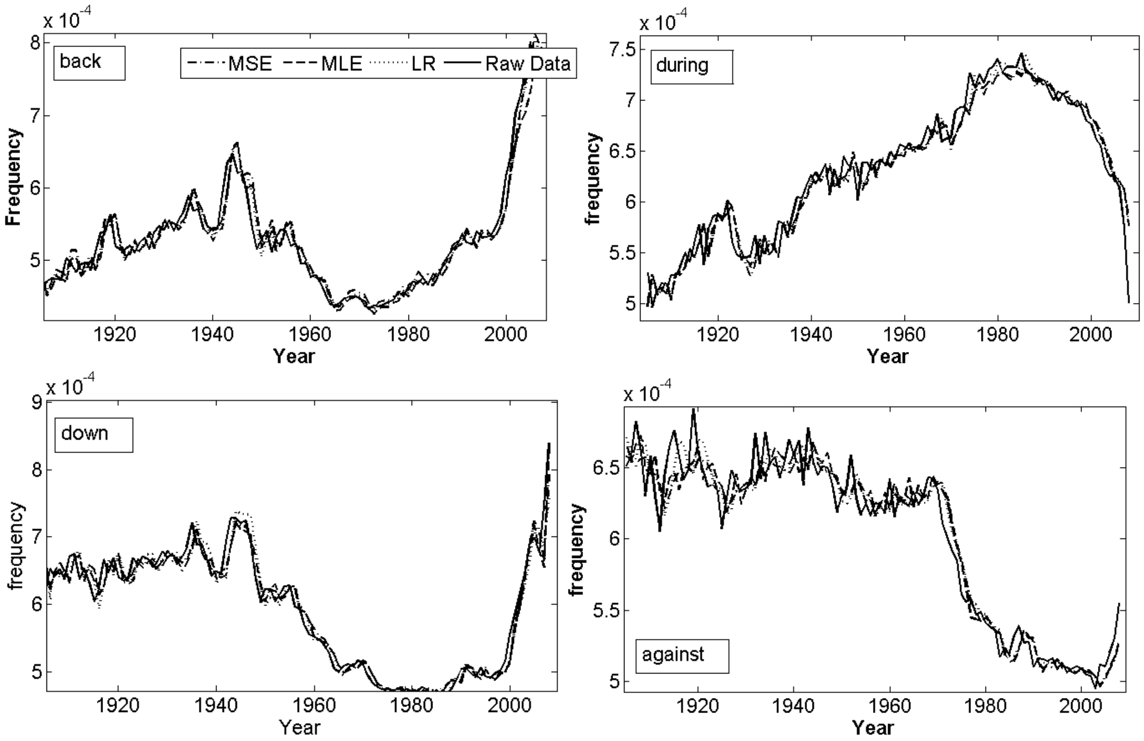

Figure 3. Predicted word usage frequency dynamics for 4 accessory words: by optimizing of MLE performance function, MSE performance function and Linear Regression model (LR).

It was shown that proposed MLE training algorithm for a neural networks is more efficient than optimization of MSE performance function for infrequent accessory words. When word usage frequency is high then the Poisson distribution is near to the Normal distribution, and the results of both methods become conformable.

This work was supported by the Russian Foundation for Basic Research (grant №12-06-00404a).

References

[1] Baayen, R.H.: Word Frequency Distributions. Springer, Berlin (2002)

[2] Gell-Mann, M., Ruhlen, M.: The origin and evolution of word order. Proc. Natl. Acad. Sci. 108, 17290–17295 (2011)

[3] Michel, J.-B., et al.: Quantitative analysis of culture using millions of digitized books. Science 331,176–182 (2010)

[4] Eduardo G. Altmann · Zakary L. Whichard Adilson E. Motter Identifying Trends in Word Frequency Dynamics J Stat Phys (2013) 151:277–288

[5] Michel J., Shen Y., Aiden A., Veres A., Gray M., The Google Books Team, Pickett J., Hoiberg D., Clancy D., Norvig P., Orwang J., Pinker S., Nowak M., Aiden E. Quantitative Analysis of Culture Using Millions of Digitized Books. Science. 331. no. 6014 . 2011. p. 176-182

[6] Maslennikova, Y. & Bochkarev. Training Algorithm for Neuro-Fuzzy Network Based on Singular Spectrum Analysis, AWERProcedia Information Technology & Computer Science. [Online]. 2013, 3, pp 605-610.

[7] Tetko, I.V.; Livingstone, D.J.; Luik, A.I. Neural network studies. 1. Comparison of Overfitting and Overtraining, J. Chem. Inf. Comput. Sci., 1995, 35, 826-833

[8] Monson,H. Statistical Digital Signal Processing and Modeling, John Wiley & Sons, 1996.

[9] Kohonen, T. 2001. Self-Organizing Maps. Third, extended edition. Springer, Berlin.