Computer Life (CPL) ISSN: 1819-4818
DELIVERING QUALITY SCIENCE TO THE WORLD

The global language development trends based on a long short term memory neural network

Ganzhou Wu

School of science, Guangdong University of Petrochemical Technology, Maoming, China

275385973@qq.com

Abstract: On the basis of the introduction of Long Short Term Memory (LSTM) model, we use the data of the top 10 languages user number from 1960 to 2017 to establish data sets, while data sets are divided into training sets and test sets. After testing, the model has higher accuracy. The LSTM network has three network structures, namely, the input, hidden, and output layers. As much, using the data set to construct the LSTM model with 10 input layers, 10 output layers and 18 hidden layers, we predict the number of speakers in 10 languages in the next 50 years. In short, except for the basic Japanese did not change much, other languages are more or less some increase.
Keywords: language development, LSTM, neural network.

1. INTRODUCTION

The development of globalization has promoted international exchanges and cooperation and language has become a bridge of transnational exchange and interaction. A large multinational service company is investigating the opening of more global offices in order to expand its international business. This article focuses on the trends of global languages and the location of new offices.

The first part of the question calls for considering various influencing factors and establishing a model that describes the trend relationship between the number of users and the time in each language. According to the established model predicts the number of native speakers in a language in the next 50 years and the total number of users in the language change trend in the trend analysis of the current world-wide use of any one language will be the other Language replaced. Combined with the projected population migration patterns for the next 50 years, we will illustrate the distribution of the geographical locations of these languages in the same period.

The second part of the question calls for an analysis of the addresses chosen for the opening of six new international offices and the language use of offices based on the modeling of the first part. And analyze the similarities and differences between such sites in the short and long term.

The rapid development of deep learning on many tasks [1] brings hope for possibly alleviating the problem of avoiding manual feature engineering. It provides a different approach that automatically learns latent features as distributed dense vectors. Recurrent neural network (RNN) [2] and its variants long-short term memory (LSTM) have been successfully used in various sequence prediction

problems, such as general domain NER, language modeling [3] and speech recognition [4]. In this paper, we propose a Long Short Term Memory (LSTM) model. Without any external resources or hand-crafted features, our neural network method can be successfully used for this task.

2. THE FOUNDATION OF LONG SHORT TERM MEMORY NEURAL NETWORK

2.1 Background

The current distribution of languages and their users in the world are extremely uneven. There are quite a few people who use languages in small numbers, while the number of users in most languages is slightly lower. According to the survey found that a small number of languages as little as only a few hundred, dozen or even a few people in use. According to David Crystal, a well-known expert on world languages, it is conservatively estimated that up to 96% of languages are used by only 4% of the world's populations. There are approximately 6,900 languages in the world, while in half the world's population, these are basically one of the languages of Mandarin (including standard Chinese), Spanish, English, Hindi, Arabic, Bengali, Portuguese, Russian, Punjabi and Japan to be as the mother tongue. A language can be the mother tongue or the second and third languages. Driven by the economic globalization, the number of users of a certain language may increase or decrease in the impact of various factors such as the level of global economic development.

The paper is mainly intended to discuss the various factors that affect the language and predict the total number of speakers in each language over time. Question is intended to discuss the prediction of the total number of speakers of various languages over time, with some other influencing factors. The language value model established in this paper obtains the 2017 top ten languages in the world through the references given in the title and uses the data to make a columnar distribution of language values. Analyze value trends of the 2017 top 10 languages in the world to predict how the total number of languages varies over time.

This is a larger sample of predictive question to solve this problem requires the use of predictive models, as well as data. First, we assume that a country only speaks a certain way of language, and use the global population data to get the number data of ten languages in 1960-2017 years, and degrade and normalize these data. Then, we built Long Short Term Memory model to predict the number of users of the ten languages in the next 50 years and determine the number of users after 50 years.

2.2 Assumptions and Notations

The assumptions are as follows.

To simplify our problems, we make the following basic assumptions, each of which is properly justified.

The people of a country will speak the language when the language is more widely distributed in this given country.

People migrate to a nation will use the language distributed in that country.

All immigrants are permanent immigrants.

The migration of population is rational.

And the notations are presented as follows.

Table 1 the notations

Number	Symbol	Significance
1	Q	The communicative value of language
3	p_i	Language popularity
4	c_i	Language centrality
5	P_i	The total number of languages spoken in all languages worldwide
6	N^s	The total number of people in all languages worldwide
7	M^s	The total number of people who speak a second language worldwide
8	h_t	The output of the hidden layer at time t
9	W_{xh}	The corresponding weight matrices from input layer to hidden layer
10	W_{hh}	The weight matrices from hidden layer to hidden layer
11	b_h	The bias for the hidden layer
12	σ_h	Activation function
13	y_t	The predict label of t-th sequence
14	W_{ho}	The weight matrices from hidden layer to output
15	b_o	The bias for the output
16	σ_y	Activation function
17	i_t	Input gate
18	f_t	Forget gate
19	o_t	Output gate
20	c_t	The memory cell of time t
21	h_t	The output of hidden layer

2.3 The model of Long Short Term Memory (LSTM)

The value of language refers to the communicative value of language, which is composed of the centrality of language and the popularity of language. For users of world languages, the communicative value of language can be expressed as [5]:

$$Q_i = p_i \times c_i = \left(\frac{P_i}{N^s}\right) \times \left(\frac{C_i}{M^s}\right) \tag{1}$$

The P_i represents the popularity of a language, which equals the total number of languages spoken in the world divided by the total number of languages spoken in all languages in the world. And c_i represents the centrality of a language, which is equal to the total number of languages in the world in which the second language is used divided by the second language in the world .The total number of people, it said the language and the global language of other languages link. This value is the

language in the global language group status. Language value is a measure of a language communicative value and language vitality indicators. The higher the language value, the higher the status of language. That is, the higher the social value, the stronger the social competitiveness, and the more people who learn the language.

The characteristic of LSTM makes it easy to model temporal sequences [6]. For the video sequences, current output depends on the current input and the previous status. More generally, suppose given input sequences denoted by $x = \{x_1, x_2, \dots, x_t, \dots, x_T\}$, where t represents th frame, and there are totally T frames. We get the formulation as following:

$$h_t = \sigma_h(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \tag{2}$$

where h_t denotes the output of the hidden layer at time t, W_{xh} stands for the corresponding weight matrices from input layer to hidden layer, W_{hh} is the weight matrices from hidden layer to hidden layer, b_h is the bias for the hidden layer, and σ_h is activation function. Finally, we can get the output through following formulation:

$$y_t = \sigma_y(W_{ho}h_t + b_o) \tag{3}$$

where y_t denotes the predict label of t-th sequence, W_{ho} stands for the weight matrices from hidden layer to output, b_o is the bias for the output, $\sigma_y$denotes the activation function.

The major problem of RNN is that it can only model the short time sequences because the error gradients vanish quickly as the networks become deeper. To solve this problem, LSTM introduces three gates to keep the status [7]. As in figure 1, there are 3 gates, including input gate (i_t), forget gate (f_t), and output gate (o_t). Where i_t and o_t control information that flows in or out the network, f_t controls the influence of previous sequences. Details are formulated as follows:

$$\begin{cases} i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \\ f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \\ o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_{t-1} + b_o) \\ c_t = f_t \odot c_{t-1} i_t + \tanh \odot (W_{xc}x_t + W_{hc}h_{t-1} + b_c) \\ h_t = o_t \odot tach\, c_t \end{cases} \tag{4}$$

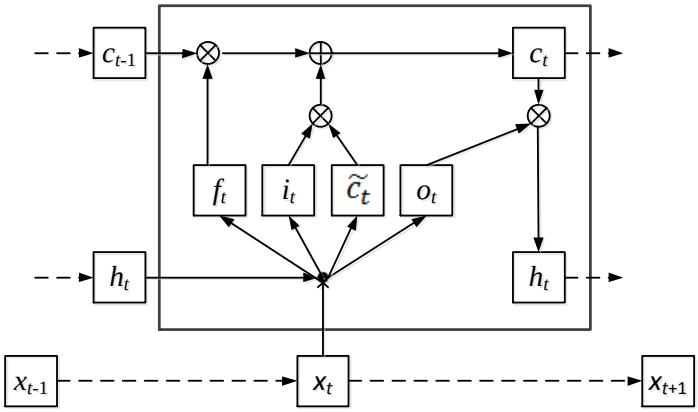

Fig. 1 The LSTM architecture

As much, using the data set to construct the LSTM model with 10 input layers, 10 output layers and 18 hidden layers, we predict the number of speakers in 10 languages in the next 50 years.

3. SIMULATION RESULTS

With the development of science and information technology, the popularity of languages is closely linked with the languages used and promoted by governments, the languages used in schools, social

pressures, the centrality of languages and the development of global tourism, international business, immigration, and electronic communications And the use of social media, the use of technology to facilitate fast and easy translation of the language is crucial. With the development of globalization, information and integration, the popularity of language is more closely connected with the centrality of language. Access to literature shows that language popularity and language center are two core factors of the language value model.

For statistical convenience, we only review the top ten languages in the list. Through the language value model to calculate the value of 2017 world's top ten languages into the software to make the world's top ten languages values in 2017 columnar distribution is reproduced in figure 2 below.

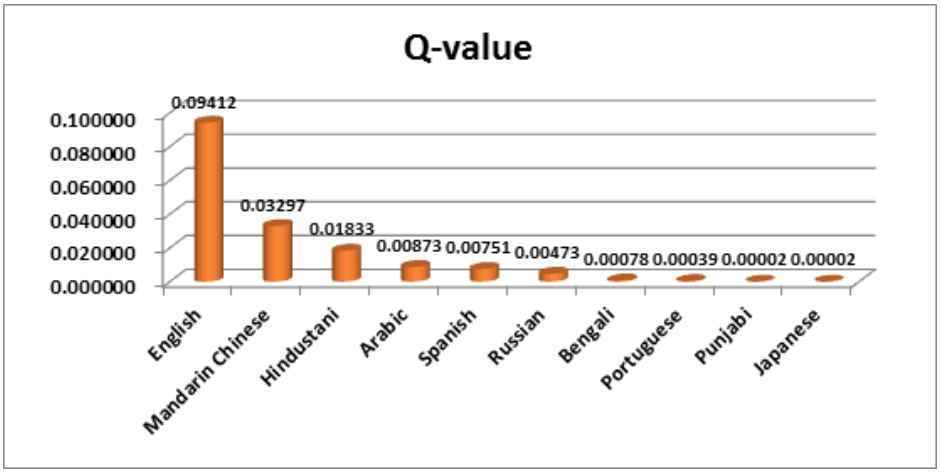

Fig. 2 Top ten languages in columnar distribution of language values

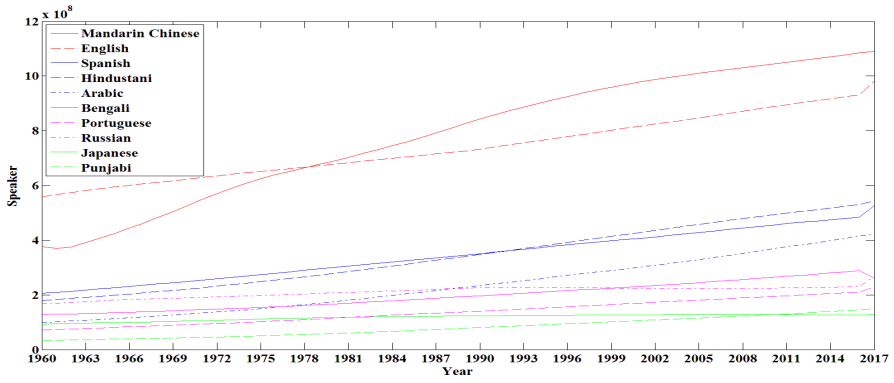

Fig. 3 Number of speakers in 10 languages between 1960 and 2017

As can be observed in the histogram, the highest value of English is 0.09412. The lowest value of Punjabi and Japanese is 0.00002. The language values of English, Putonghua and Hindi are much higher than those of the other seven languages, and the vitality of the language is strong. We can be anticipated that over time, the number of speakers of English Putonghua, and Hindi will have the number of speakers of English is comparatively increased. Slightly higher values of Arabic, Spanish and Russian languages suggest that the number of Arabic, Spanish and Russian languages will hardly change much in the future. The low values of Bengali, Portuguese, Punjabi and Japanese are at a disadvantage in the top ten mothers in the world, which will reduce the number of speakers in Bengali, Portuguese, Punjabi and Japanese, where The relative decrease in Punjabi and Japanese is most pronounced.

We assumed that a certain country speaks only one language and screened in the world for countries that speak the top ten languages, and obtain data on the number of people using the 10 languages between 1960 and 2017. The datum is shown in figure 3 below.

There is a trend of data, so first eliminate the data trend, and then the data normalized, and the first two thirds of the dataset as a training set, the latter as a test set to build the input node is 10, the output node 10, hidden nodes for the 18 time-recurrent neural network, the use of test data to test the reliability of the model.

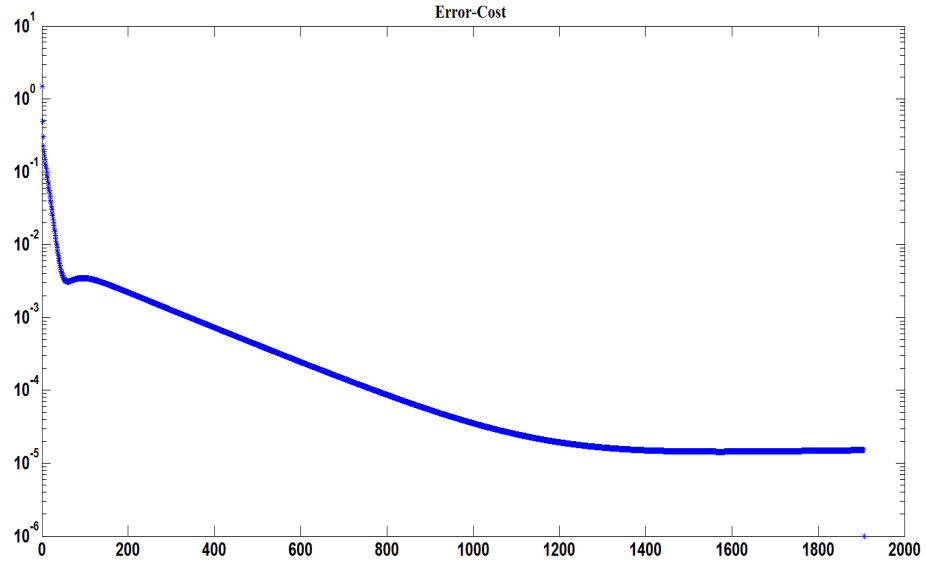

Fig. 4 Error-Cost

From the figure4, we can see that the number of iterations around 1400 convergence, the error tends to 0.00001 or so, indicating that the model is based. After testing, the model is more reliable. Therefore, this paper builds the following parameters time-based recurrent neural network model to predict the total number of ten language speakers in the ensuing 50 years. The number of iterations is placed at 2000, the activation function uses the sigmoid function.

The forecast results are shown in figure 5 below:

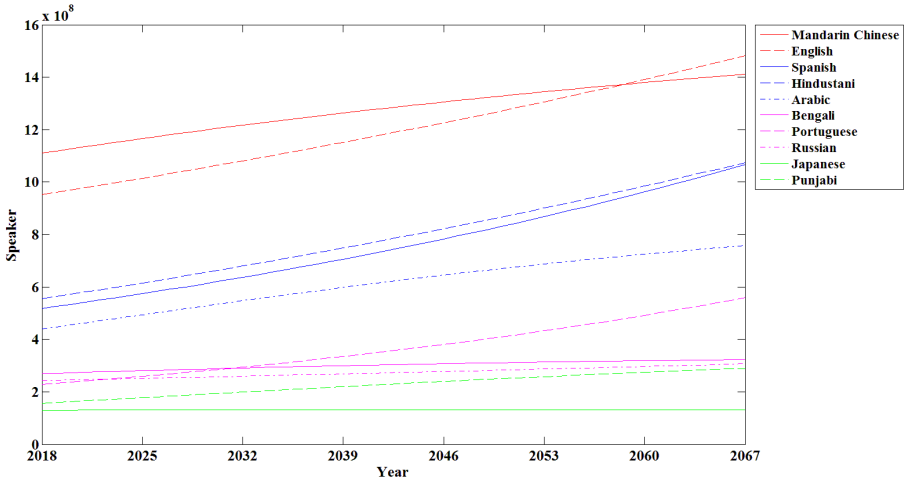

Fig. 5 The forecast results

As can be observed in the figure5, the number of language users in Putonghua, English, Spanish and Hindi has risen dramatically over the next 50 years. In Punjabi, the total number of Russian speakers is relatively low after 50 years. The number of Japanese speakers will not increase significantly in the next 50 years.

4. CONCLUSION

We can see that the Q value of the four languages of Mandarin, English, Spanish and Hindi are ranked in the top four of the ten languages, while the Q values of Russian, Punjabi and Japanese are the last. The greater the Q value, the stronger the viability of the language, so Japanese, Russian, Punjabi in the future is liable to be replaced by other languages. On the basis of the introduction of Long Short Term Memory (LSTM) model, we use the data of the top 10 languages user number from 1960 to 2017 to establish data sets, while data sets are divided into training sets and test sets. After testing, the model has higher accuracy. Finally we predict the number of speakers in 10 languages in the next 50 years. In short, except for the basic Japanese did not change much, other languages are more or less some increase.

ACKNOWLEDGEMENTS

This paper was supported by National Natural Science Foundation of China Youth Fund Project (NO.11501131), Guangdong Youth Innovative Talent Project (NO.2014KQNCX202), Guangdong Province, the outstanding young teachers training program funded projects (NO.YQ2015117).

REFERENCES

[1] Zeng T, Li R, Mukkamala R, Ye J, Ji S, “Deep convolutional neural networks for annotating gene expression patterns in the mouse brain”, BMC Bioinformatics, 2015,Vol.16(1),p1-8

[2] Chiu JP, Nichols E, “Named entity recognition with bidirectional lstm-cnns”, Trans Assoc Comput Linguist, 2016,Vol. 38(4),p357-370

[3] Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J, “Distributed representations of words and phrases and their compositionality”, In advances in Neural Information Processing Systems.,2013, Vol.25(3),p111-119

[4] Graves A, Jaitly N. Towards end-to-end speech recognition with recurrent neural networks. In: Proceedings of ICML 2014,Beijing: International Machine Learning Society,2014.

[5] SU Jian, “Q Value and Survival Region of Minority Languages”, A Model Based on Language Economics,2011, Vol.21(1),p53-54.

[6] GE Rui, “Action recognition with hierarchical convolutional neural networks features and bi-directional long short-term memory model”,Computer Simulation. 2017, Vol.34(6),p79-86.

[7] Guan N, Tao D, Luo Z, et al, “Online nonnegative matrix factorization with robust stochastic approximation”, IEEE Transactions on Neural Networks and Learning Systems, 2012, Vol.23(7) ,p1087–1099.