Morphological text analysis using neural networks

A.N. Zhdanova
zhdan.aleksandra@gmail.com

A.V. Kupriyanov
akupr@ssau.ru

D.S. Sherenkov
dsherenkov000@gmail.com

TECHNOLOGY OF NEURAL NETWORKS

Recurrent neural networks successfully cope with the tasks associated with the classification of sequences, particularly with morphological marking. There are many studies in which a recurrent neural network was used to solve the problem of part-of-speech labeling. One of the goals of this work is to try to use a simpler neural network model to solve the part-string labeling problem to prove that such a model is also capable of coping with the task, not much worse than more complex models. The model of a multilayer perceptron was chosen to prove this assumption.

PROBLEMS OF NATURAL LANGUAGE PROCESSING

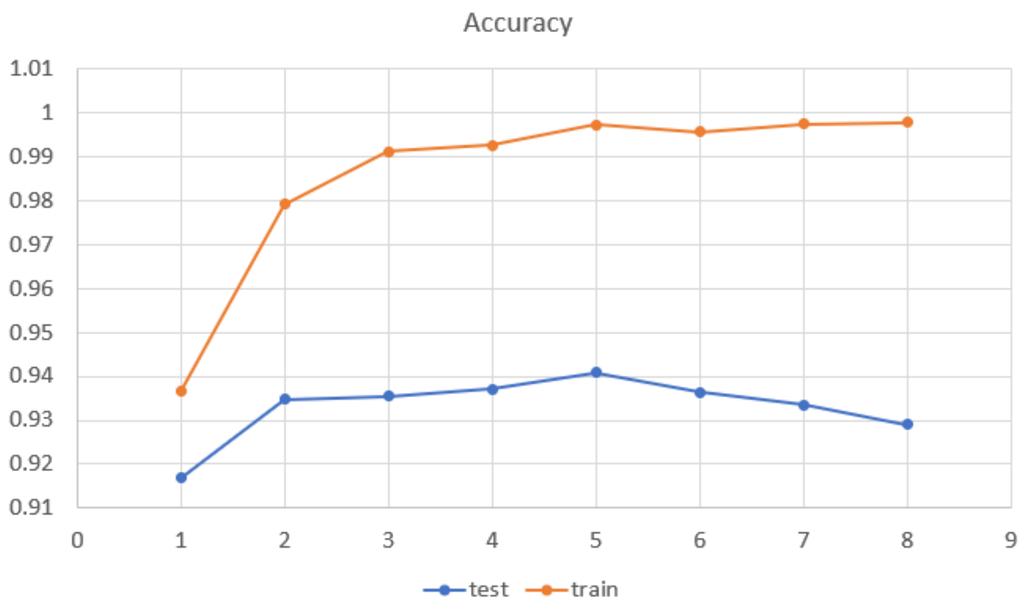
- Tokenization (separating the text into words and lexemes)
- Supply boundaries
- Sentence semantics (there may be several options for parsing some sentences)
- Suppletivism (when the forms of the same word are formed from different words)
- Homonymy

NEURAL NETWORK TRAINING

The neural network was trained on a training dataset that consists of 25,000 words that make up 1,000 sentences. Each word and punctuation mark is assigned a tag that defines a part of speech and some morphological features.

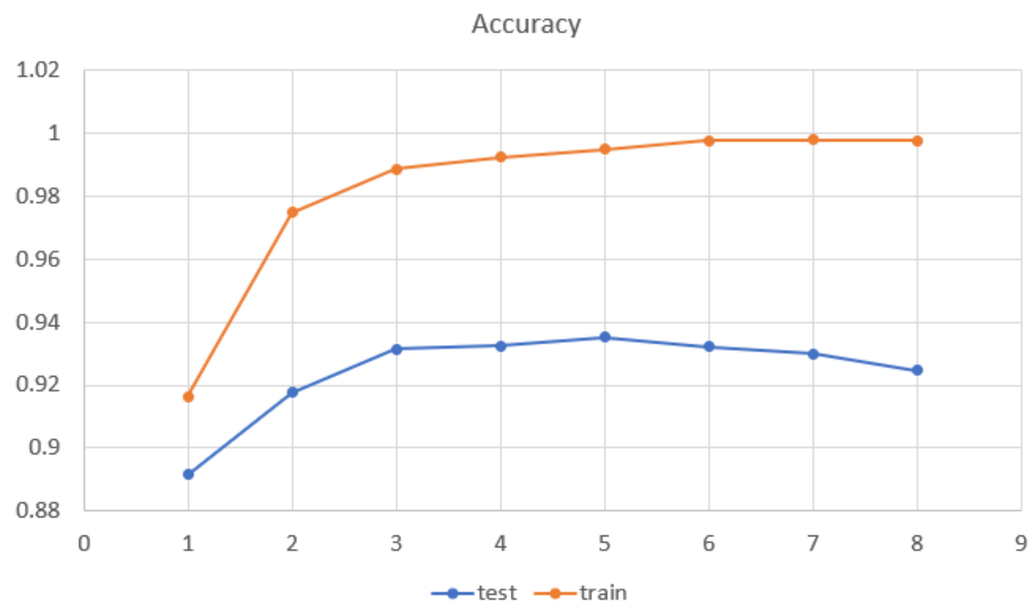
```
[('Japanese', 'ADJ'), ('investment', 'NOUN'), ('in', 'ADP'),
('Southeast', 'NOUN'), ('Asia', 'NOUN'), ('is', 'VERB'),
('propelling', 'VERB'), ('the', 'DET'), ('region', 'NOUN'),
('toward', 'ADP'), ('economic', 'ADJ'), ('integration', 'NOUN'), ('.', '.')]
```

THE ACCURACY OF THE MARKUP OF THE ENGLISH TEXT



Graphs of dependences of the accuracy of the markup of the English text on the number of epochs

THE ACCURACY OF THE MARKUP OF THE RUSSIAN TEXT



Graphs of the dependence of the accuracy of markup of the Russian-language text on the number of epochs

RESULTS

Numb. of epochs	Training set English labeling accuracy	Test set English labeling accuracy	Training set Russian labeling accuracy	Test set Russian labeling accuracy
1	0.9365	0.9167	0.9165	0.8916
2	0.9792	0.9346	0.9749	0.9176
3	0.9911	0.9354	0.9888	0.9315
4	0.9927	0.9291	0.9925	0.9324
5	0.9973	0.9407	0.9953	0.9352
6	0.9957	0.9363	0.9976	0.9322
7	0.9974	0.9335	0.9980	0.9298
8	0.9978	0.9289	0.9978	0.9246

CONCLUSION

The maximum markup accuracy for the test dataset in both cases is achieved after five training epochs and is 94% and 93.5% for English and Russian texts, respectively. This model copes with English text better than Russian. This can be explained by the fact that the English language has stricter rules for the placement of words in a sentence, which means that in the Russian text, it is more difficult to rely on the context when determining the part of speech of a word.

SAMARA UNIVERSITY

34 Moskovskoye shosse, Samara, 443086
+7 (846) 335-18-26, +7 (846) 335-18-36 (fax)
ssau@ssau.ru, www.ssau.ru

Faculty of Information Technology, Software Department
+7 (846) 267-43-61 , +7 (905) 302-39-58
zhdan.aleksandra@gmail.com