The complex of neural networks and probabilistic methods for
mathematical modeling of the syntactic structure of a sentence of natural language

A Sboev, R Rybka, I Moloshnikov, D Gudovskikh
NRC “Kurchatov Institute” (Academic Kurachatov sq. 1, Moscow), Department of
mathematical modeling and information technologies, Russia

E-mail: sag111@mail.ru, RybkaRB@gmail.com

Abstract. The formalized model to construct the syntactic structure of sentences of a natural
language is presented. On base of this model the complex algorithm with use of neural
networks founded on data of Russian National language Corpus and set of parameters extracted
from this data was developed. The resulted accuracy along with possible accuracy which
theoretically could be received with these parameters is presented.

1. Introduction
The growing intensity of information exchange in the virtual environment, in particular in social
networks, makes it necessary to use mathematical modeling techniques and complex software for
automated analysis of data from the virtual environment, and for displaying social phenomena in them.
In this area, typical application tasks are: automated annotation of texts, content analysis of business
information, sentiment analysis, an analysis of emotive text, identifying threats in social networks. To
solve these problems, the computerization of the main stages of the analysis of unstructured data is
crucial. In particular, it refers to a problem of determining the syntactic structure of the text, written in
natural language. It is a key addition to the arsenal of BigData technologies for texts analysis. The
classic approach to automating analysis of objects and phenomena is their mathematical modeling.
Syntactic Parsing is performed by linguists on the basis of the natural language rules, for which it is
not possible to fully build a rigorous mathematical theory. Existing solutions for obtaining the
syntactic structure of a sentence are based on two main approaches: the first includes the software
implementation of linguistic parsing rules, and the second includes mathematical algorithms with pre-
training on a set of examples with syntactic markup. The first are not universal and poorly tolerated in
other languages, and time-consuming in implementation (such as the ETAP-3 for the Russian
language has been being created more than 20 years). Syntactic analysis methods based on algorithms
with pre training include: systems based on a combination of parsing algorithms such as CKY or
EARLY, as well as their modifications, with probabilistic grammars (Link Grammar, PCFG); neural
network method based on a combination of RAAM and SRN; algorithms based on combination of
transition systems and pre-trained classifier. Application of the above techniques in conjunction with
the parametric description of the words of sentence has its peculiarities. In particular, the methods of
formal grammars are mainly used for languages without direct order of words and with projective
connections. Parsers based on recurrent neural networks [1] lose their analysis accuracy with
increasing the number of words in a sentence. On the whole, the approach based on neural network
models has some advantages because neural networks have some generalizing properties. The authors of [2] [3] have shown that application of convolution neural networks to extract features of sentence in conjunction with methods of deep learning [4] contributes to increase in quality of the shallow parsing and determination of parts of speech when analyzing sentences of the English language. The specialists at Stanford University proposed a method based on the replacement of words and parts of speech by their probability vectors formed based on the Word2Vec model [5] that is trained on a corpus of texts. They created parsers for some languages (English, Chinese, German, Arabic) [6], which are based on this idea. Parsers based on models of transitions [7] are widely used at present, but a full-fledged research of classification models in its composition is not submitted.

This paper proposes a combined approach to the creation of a mathematical model of the syntactic structure of the sentence, which involves:

- mathematical formalization of linguistic procedure of syntactic analysis;
- development of methodology of parameterization of words of sentence for choosing the parameters allowing to establish syntactic relations with minimal ambiguity;
- assessment of the achievable accuracy of syntactic parsing on the basis of used corpus dataset;
- selection of methods for determining syntactic relations and formation of syntactic structure of sentence.

In this study the base "SinTagRus" from the Russian National Corpus [8] is used for training neural network models in the composition of a model of syntactic structure of sentence. It contains sentences with unique morphological and syntactic markup. Examples from the RNC can be regarded as a space of pairs of words that include syntactic relations between words with markings of the relations, the word in its composition, and supplementary information of sentence. In this space, when using neural network methods, actually a measure is created that can be used to assess the syntactic proximity. This leads to application of probabilistic methods to eliminate ambiguities arising after parsing.

2. Material and methods

2.1. Mathematical formalization of linguistic procedure of syntactic parsing

The syntactic structure of natural language includes structures of individual sentences. Syntactic structure of the sentence $F = \{f_1 ... f_n\}$ (where $f_i$ is word in this sentence) can be formalized in a tree with oriented edges (arcs) $T(W,D_T)$. Here $W$ is set of vertices $w_i = \{f_i,l_i,m_{li}\}$. The linguistic characteristics of $w_i$ are:

- $l_i$ – the normal form of the word (lemma);
- $m_{li}$- Morphological characteristics: the part of speech, gender, number, case, etc;
- $D_T$ – a plurality of oriented arcs in the parse tree. In the tree, the number of arcs $N_D = n - 1$, where $n$ is the number of words in the considered sentence.

Thus, linguistic parsing procedure of sentence can be formalized as a sequence of graphs $G_i$, which should end with the construction of the tree: $G_0 \Rightarrow G_1 \Rightarrow \cdots \Rightarrow G_{i-1} \Rightarrow T$. According to the presented formalization, we described an automatic procedure for determining of syntactic relations and for building of syntactic structure of sentences (Syntactic parsing tree) in general forms:

1) subgraph $g$ that consists of 2 vertices and their environment is selected from the graph $G_i$;
2) subgraph $g$ analyzed by formal neural networks for establishing syntactic relation between 2 selected vertices;
3) if neural network analysis of subgraph $g$ has shown that syntactic link is established between the two selected vertices, then the new arc is added to the graph $G_{i+1}$, and a new graph $G_{i+1}$ is created. If the establishment of syntactic relation does not happen, then the other subgraph $g$ is selected from the graph $G_{i}$.

In the event that all vertices that have no relations from examined vertices of graph, and all possible subgraphs $g$ in the graph $G_i$ have been enumerated, but the addition of a new arc does not happen, then following options are available depending on the algorithm of busting subgraphs $g$ from $G_i$:
1) $G_i$ is a tree. Then parsing of sentence is completed;
2) $G_i$ can be represented in the form of several trees $T_k(W,D_T)$. Then one of them, which is accepted as the final result of the parsing of sentence, is selected using probabilistic methods;
3) $G_i$ cannot be decomposed into a set of several trees. In this case, the syntactic structure of sentences is not determined.

### 2.2. Parameterization methodology

The purpose of this technique is the choice of parameterization of words of sentence to form an input vector for a neural network. Criterion for selection of a particular parameterization is based on the number of clear determined syntactic relations by the statistical analysis of the texts of RNC. The parameterization was held on four groups of parameters [9] (table 1). They contain common parameters (morphological characteristics of words; punctuation signs and indicator of capital letter) and parameters supplemented by us. The latter includes parameters of potential syntactic relationships ($p_{sinto}$) that are established on the basis of morphological characters of two words. $p_{sinto}$ indicates the possible syntactic connections within a sentence. Thus, efficacy was evaluated by the number of ambiguous relationships: the higher is the number of ambiguous relations, the worse is the efficacy of a set of parameters.

#### Table 1. Description of sets of parameters

| Parameters \ set № | 1 | 2 | 3 | 4 |
|-------------------|---|---|---|---|
| Morphological characters | + | + | + | + |
| Additional | + | + |
| The displacement of the main word to the dependent word of the pair in sentence | + | + |
| Potential syntactic relations ($p_{sinto}$) between main and dependent words (pair of words) | + | + |
| Potential syntactic relations from the words of the pair to other words in expression | + |

### 2.3. Methods of syntactic parsing

#### 2.3.1. Approaches to building a syntactic parse tree and to formation of training examples to determine syntactic relations

Two approaches were investigated to construct a syntactic parse tree: the first one is based on exhaustive enumerating all possible options for establishing Sinto between words in a sentence; the second one is based on the Covington scheme [7] of incremental parsing. In the first case, training set consists of pairs of words from all RNC sentences divided into two classes: 1) form Sinto; 2) do not form a Sinto. The number of examples in the second class is much larger than in the first. Therefore we build a method of filtering the pairs of words that form sinto. After that we determine syntactic relations in the filtered set of examples.

The essence of the second approach lies in building a model of transitions in three lists of words: the Right list consisting of all unparsed words, the Left list comprising the word for finding relationship between the heads of Right and Left lists, and the intermediate list. If relation between head of Right and Left lists has not been found, then the intermediate list will be replenished with the word from the head of the Left list. Four classes of actions (No-Arc, Shift, Right-arc, Left-arc) are used for relocation of words between the lists and the definition of syntactic context. In this case, the training set consists of examples corresponding to these actions according to the type of sinto.

#### 2.3.2. Determining syntactic relations

The following solutions are used for classification of Sinto and action of transitions:

1) Creating a sequence of classification neural network models to determine a Sinto or action independently (binary classification). The sequence is formed based on the number of examples for each class: from biggest to smallest;
2) Combining Sinto into several groups based on the number of examples for them, and on the accuracy of models for an independent classification of syntactic relations (or actions). Neural network models are created for each group of Sinto;

3) Creating a single model for multiclass classification for all Sinto (or actions).

In the first and second cases, each next classifier is trained on base of a training set which is free from the examples used for learning previous models. For constructing classification model, several methods where investigated. The main goal of them is to determine the syntactic relations of sentence. We explored a large variety of methods for improving the accuracy of the classification, including AdaBoost, Boosting by filtering, for reducing the input space of attributes, including PCA, ICA, RBFsampler. Methods MLP, SGD [10], SVM strategy one-vs-all [11], PNN, GNT [12], ensembles of decision trees (RFC) [13] in combination with the methods of reducing the dimension of the input space (Nystroem) [14] were also investigated and demonstrated high scores (see section Experiments). For estimates of the models we used indicators of Positive and negative predictive values (PPV and NPV).

3. Experiments.

3.1. Parameterization of words of sentence

The results of Table 2 show that the fourth set of parameters has the best effectiveness. Therefore, the fourth set of parameters will be used for subsequent parameterization of words of the sentence in the input vector for the classifier based on neural network. After selecting the parameterization of words, the RNC data is analyzed to assess the achievable accuracy of syntactic parsing. Syntactic trees in RNC have several properties:

- the parse tree has only one vertex;
- there is only one input connection for all words in a sentence except for vertex;
- syntactic parse tree includes all the words in the sentence.

| № set of parameters | Average number of ambiguous syntactic relations for word of sentences | Percentage of clear syntactic relations determined |
|---------------------|--------------------------------------------------|-----------------------------------------------|
| 1                   | 102.29                                           | 58.48                                        |
| 2                   | 56.28                                            | 78.9                                         |
| 3                   | 8.84                                             | 85.72                                        |
| 4                   | 1.43                                             | 98.91                                        |

Table 2. Comparing sets of parameters

From this perspective, we have formed the criteria for the effectiveness of a selected set of parameters to syntactic parsing:

- the number of sentences with unambiguous parsing;
- the number of sentences with ambiguous parsing;
- the average number of syntactic trees for ambiguous-parsed sentences.

The evaluation results show that the proportion of uniquely-parsed sentences is 79.9% of the total number of proposals (42.9 thousand). The average number of parses of ambiguous-parsed sentences (20.1%) is 21.3. Classifications based on neural network PNN, MLP and SVM are designed to establish parameter of p_sinto. Average of PPV, NPV of setting p_sinto are very high (PPV=99.9%, NPV=99.8%)

3.2. Building procedure of syntactic parsing

For determine syntactic relation we investigate several approach and classification methods (see chapter 2.3). The best results demonstrate approach based on incremental parsing scheme (Table 3). In this case the training sample is about 1.35 million examples.
Table 3. PPV and NPV for defining actions in incremental parsing scheme. Multiclass is model based on ensembles of decision trees or SVM with strategies such as One-vs-all. Binary is the decision on the basis of a binary classification action when a classifier is constructed for each action.

| Variant of descent | Using methods           | PPV  | NPV  |
|--------------------|-------------------------|------|------|
| multiclass         | SVM (linear kernel)     | 90.1 | 91.2 |
| multiclass         | Nystroem + RFC          | 83.8 | 84.1 |
| binary             | SVM + SGD               | 87.3 | 88.1 |
| binary             | Nystroem + RFC and SVM + SGD | 89.2 | 90.1 |

Thus, the approach based on combination of the incremental parsing scheme and the classifier based on SVM with linear kernel was selected for the implementation of the methods of modelling syntactic structure of Russian language sentence.

3.3. Testing
The results of testing the implemented model and comparison with other systems are presented in Table 4.

Table 4. Estimation of syntactic parsing and comparison with estimation from literature sources “Estimation 1” obtained by using method presented on previous section. For obtaining “Estimation 2” we add word forms in set of parameters. Estimation from literature sources 1 is based on the use of ETAP-3 [15]. Estimation from literature sources 2 is presented in [16]. UAS – unlabeled attachment score; LAS – labeled attachment score; TRD – is the ratio of number sentences with correct root determination to number of sentences; TSSP – is the ratio of number sentences with right structure of syntactic parse tree to the total number of sentences; TSPT – is the ratio of number sentences with right syntactic parsing tree to the total number of sentences.

| Task description                  | UAS (%) | TRD (%) | LAS (%) | TSPT (%) | TSSP (%) |
|-----------------------------------|---------|---------|---------|----------|---------|
| Our estimations 1                 | 85.81   | 82.23   | 79.33   | 14.05    | 29.47   |
| Our estimations 2                 | 91.73   | 88.84   | 89.39   | 35.91    | 52.38   |
| Estimation from literature sources 1 | 94.3      | ---    | 92.3    | 29.7     | 37.4    |
| Estimation from literature sources 2 | 89.4      | ---    | 83.4    | 21.8     | 33.3    |

Testing of models to determine syntactic relations on the corpus sentences unused during training showed that the accuracy of determining the type of syntactic relations is equal to 79.33%, and the accuracy of the forming syntactic tree is equal to 14.05%. Adding word forms to the selected set of parameters increases the accuracy of the determination of syntactic relations by 10% and the one of forming syntactic tree by 20.7%.

4. Conclusion
A combined approach based on proposed formalization of linguistic procedure to modeling of syntactic structure of a sentence in Russian language was presented in this paper. Set of combinations of methods of constructing a syntactic structure, of establishment of syntactic relations, and parameterization of words were investigated. The best accuracy of syntactic parsing was obtained by using a combination of: selected parameterization of the words, including the potential syntactic relations; the neural network methods for their establishment; the classifier based on support vector machines with linear kernel and incremental parsing scheme. In general, the results of testing of the syntactic parsing algorithm on the basis of the developed model are higher than ones from the literature sources (see [17], [16]). Estimates of the achievable accuracy of parsing have shown that the quality of the parsing can be significantly improved by using a more accurate algorithm for determining the syntactic relations. In further, our work will focus on improving methods of establishing syntactic relations in a sentence.
5. Acknowledgements
Numerical simulations were performed at supercomputing resources in NRC "Kurchatov Institute", which are supported as the centre for collective usage (project RFMEFI62114X0006, funded by Ministry of Science and Education of Russia).

6. References

[1] Wong Chun Kit, “Recursive Auto-Associative Memory as Connectionist Language Processing Model-Training Improvements via Hybrid Neural-Genetic Schemata,” City University of Hong Kong, Hong Kong, 2004.

[2] R. Collobert, “AISTATS,” in Deep Learning for Efficient Discriminative Parsing, 2011.

[3] Bengio Y. et al., “A neural probabilistic language model,” The Journal of Machine Learning Research, vol. 3, pp. 1137-1155, 2003.

[4] LeCun Y., Bengio Y., Hinton G., “Deep learning,” Nature 521, p. 436–444, 2015.

[5] “Tool for computing continuous distributed representations of words.” Google inc., [Online]. Available: https://code.google.com/p/word2vec/.

[6] Chen D., Manning C.D., “A fast and accurate dependency parser using neural networks,” Proc. EMNLP, p. 740–750, 2014.

[7] Nivre J., “Incrementality in Deterministic Dependency Parsing,” School of Mathematics and Systems Engineering, Vaxjo, Sweden, 2004.

[8] "Russian National Corpus," [Online]. Available: http://www.ruscorpora.ru/instruction-syntax.html.

[9] Rybka R. et al., “Statistically selected set of parameters for definitions of syntactic relations in Russian language sentences,” System analysis and information technologies:Journal of the Voronezh State University., no. 2, pp. 117-124, 2014.

[10] Zhang T., “Solving Large Scale Linear Prediction Problems Using Stochastic Gradient Descent Algorithms,” in ICML 2004, 2004.

[11] Rocha A., Goldenstein S., “Multiclass from Binary: Expanding One-vs-All, One-vs-One and ECOC-based Approaches,” in IEEE Transactions on Neural Networks and Learning Systems, 2013.

[12] Sboev A., et al, “Neuronetwork package Neurotree adapted for the segment of the Russian grid network,” Informatization and Communication, 2012.

[13] L. Breiman, “Random forest,” University of California, Berkeley, 2001.

[14] Kumar S., Mohri M., “ Ensemble Nyström Method,” Courant Institute of Mathematical Sciences, New York, 2009.

[15] Jurij Apresjan et al., “ETAP-3 Linguistic Processor: a Full-Fledged NLP Implementation of the Meaning <->Text Theory,” Laboratory of Computational Linguistics Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow.

[16] Sharoff S., Nivre J., "The proper place of men and machines in language technology Processing Russian without any linguistic knowledge," in Proc. Dialogue 2011, Russian Conference on Computational Linguistics., Moscow, 2011.

[17] Kazennikov A., "A comparative analysis of machine learning," Papers from the Annual International Conference “Dialogue”, no. 9 (16), pp. 157-163, 2010.

[18] Kübler S. et al., Dependency Parsing Synthesis Lectures on Human Language Technologies, 2009.