Procedia Computer Science

Volume 66, 2015, Pages 277–286

YSC 2015. 4th International Young Scientists Conference on Computational Science

Syntactic analysis of the sentences of the Russian language based on neural networks

A.G. Sboev[1, 2], R. Rybka[1], I. Moloshnikov[1], and D. Gudovskih[1]

[1]NRC "Kurchatov Institute", Moscow

[2] MEPhI National Research Nuclear University, Moscow

sag111@mail.ru, RybkaRB@gmail.com, ivan-rus@yandex.ru, dmitrygagus@gmail.com

Abstract

The model of Russian language parser based on a combination of neural networks along with extraction of set of parameters which allows to establish relations with the minimal syntactic ambiguity is presented. The parse tree of sentence is constructed in the format of Russian National Corpus (RNC). RNC texts containing morphological and syntactic markup are used for training neural network models as part of procedure. Estimates of accuracy of the developed parser procedure in comparison with the other Russian language parser systems have been performed.

Keywords: neural networks, natural language processing, syntactic analysis, dependency parsing

1 Introduction

Currently increasing intensity of information exchange leads to the necessity to develop automated systems for text processing for the annotation of documents, content analysis of business information, sentiment analysis, emotion analysis of the text, threats identification in the social networks, etc. A key aspect of the quality of such systems is the way to establish relations between words in a separate sentence.

The past few years, language corpuses have been forming actively. The labeled examples from them allow to determine relations between words using statistical methods and techniques of artificial intelligence in particular neural networks. This approach seems more universal than to formalize a great set of rules and regulations for a particular language. For the Russian language, the task is complicated by the presence of free order of words, non-projective relations in the sentence, and complex morphology (for example, a system of such type "ETAP-3" (Iomdin L. et al., 2012) has been being developed more than 20 years). Moreover, an adaptation of the system of this type for other languages is often simply not possible.

Selection and peer-review under responsibility of the Scientific Programme Committee of YSC 2015

doi:10.1016/j.procs.2015.11.033

The aforesaid actualizes the task of developing an algorithm to establish syntactic relations (here called sinto) between the words and the formation of the syntactic parse tree of sentence based on data from the National Language Corpus.

An effective solution of this task with using methods of corpus linguistic necessitates the complex study, including, on one hand, what accuracy of parsing can be achieved with the data from Language corpuses, and on the other hand, which methods can be used to achieve such accuracy.

Currently various methods are used to establish the syntactic relations, in particular, the methods of probabilistic grammars (PCFG, Link Grammar), methods of artificial neural networks (SVM, SRN, and RAAM, and others).

At the same time different types of corpus information are used:

- grammatical characteristics of individual words with the addition of features that characterize the properties of writing words, the presence of separators, their place in the sentence, and so on;
- signs of an individual word form of the corpus dictionary.

In all cases, the application of given or other combinations of methods and information of corpus has its disadvantages and advantages for specific language. In particular, the methods of formal grammars are mainly used for languages with projective connections, while parsers based on recurrent neural networks (Wong Chun Kit, 2004) lose their accuracy analysis increasing the number of words in a sentence. On the whole, the approach based on neural network models has some advantages due to the fact that neural networks exhibit known generalizing properties. Using these properties combined with parametric description of words and modern methods of data compression helps to reduce the dimensionality of the address space of the attributes of the task and to build techniques universal for different languages.

In (Collobert R. et al., 2011) the authors consider the method of forming features with classification convolution neural networks (LeCun Y. et al., 2010). Extraction of features can be performed either within whole sentence, or within its window of words. Construction of structure of the sentence is done using HMM. As shown in the article (R. Collobert. , 2011), approach based on deep learning (LeCun Y., Bengio Y.,Hinton G., 2015) has demonstrated good results for English texts in solving problems of POS-tagging and chunking on IOBES format.

An alternative approach uses a language model with features extraction of words based on the probabilities of co-occurrence of words in the training corpuses presented in the works (Bengio Y. et al., 2003). Syntactic (dependency) parser for some languages (English, Chinese, German, Arabic) (Chen D., Manning C.D., 2014) based on this approach has been built using a hybrid neural network.

Methods based on the model of transition (Kübler S. et al., 2009) (Sharoff S., Nivre J., 2011) (Nivre J., 2004), use a combination of features in a context of words in a sentence, and information on the previous analysis. This type of transition (or parsing rules) for the current state of analysis is calculated using the SVM.

In this paper, the model of parser based on a combination of neural networks along with extraction of set of probability parameters, which allows to establish relations with the minimal syntactic ambiguity, is presented.

Its main function is syntactic parsing of sentence based on the format of used language corpus. The basis for the study are sentences with unambiguous morpho-syntactic marking from (Russian National Corpus) (RNC). We investigate achievable accuracy on RNC markup sentences.

Further, in Chapter 2.1, the rationale of the choice of parameters is presented. It involves a study of parameter combinations that exhibit less ambiguity in determining syntactic relations. Chapters 2.2 and 2.3 contain a description of approaches for the establishment of syntactic relations, and the formation of the parse tree. In Chapter 2.4, criteria for evaluating the result of parsing are presented. Section 3.1 is dedicated to experimental results. Finally, we evaluate the syntactic parsing procedure and show the further development of the algorithm for constructing the parse tree.

2 Materials and methods

2.1 Selection of the set of parameters for syntactic parsing and assessment of the accuracy of syntactic parsing on the basis of them

We have investigated four groups which consist of a set of common parameters and another set supplemented by us. The former includes: morphological characteristics of words; additional ones, such as indicator of punctuation after the word and indicator of capital letter; the distance between words. The latter includes potential syntactic relationships that are established on the basis of morphological characters of two words, further called p_sinto.

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				+

Table 1: Description of parameter sets

The effectiveness of a set of parameters was evaluated by determining the degree of ambiguity in establishing syntactic relations described by this set. For this purpose the special procedure has been created (Rybka R. et al., 2014), based on count of pairs of words in sentences from RNC and results of their parsing. Efficacy was evaluated by the number of ambiguous relationships: the higher is the number of ambiguous relationships, the worse is the efficacy of a set of parameters.

№ 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

The results of Table 2 show that the fourth set of parameters has the best effectiveness. Further we will assess the accuracy of syntactic parsing using the 4-th parameter set.

The result of syntactic parsing is the syntactic tree of a sentence. At that the nodes correspond to the words or to their characteristics. The arcs correspond to the links, and their syntactic types. In accordance with the format used by the RNC, syntactic trees have several properties:

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

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 number of RNC sentences having after parsing properties a-c is calculated. If the sentence after parsing does not have the "b" property, then the procedure of normalization is being carried out. Its goal is to transform each ambiguous parsing to several syntactic trees. If after normalization procedure the result does not have "a" and "c" properties, then this is an error of parsing.

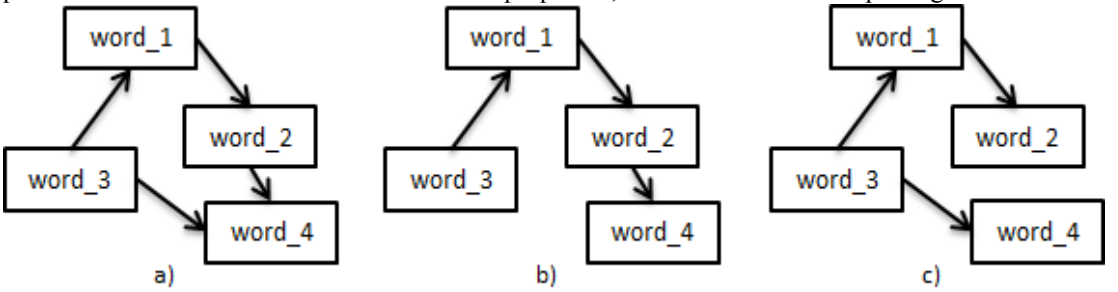

Figure 1: Example of ambiguous parsing (a) and results of their normalization (b,c)

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 p_sinto. Positive and negative predictive values (PPV, NPV) of setting p_sinto are very high (see Table. 3)

Number of syntactic relations type	Best model of NN	PPV/NPV
1,32,33,42,52,54,55,57,58	PNN	99.9/99.8
5,6	SVM	99.9/99.9
2-4, 7-31, 34-41, 43-50, 56, 63-76	MLP 2 layer (40, 20 neurons)	99.8/99.6

Table 3: Best model to determine p_sinto

In the case of a MLP neural network, here and below the number of neurons in the hidden layer was chosen using a genetic algorithm.

Thus, for future work, we have chosen a set of parameters (Rybka R. et al., 2014), including morphological characteristics, features of capitalization and punctuation, as well as p_sinto established on the basis of morphological characters of the two words in the sentence.

2.2 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 of all possible options for establishing Sinto between words in a sentence;
- the second one is based on the Covington scheme (Nivre J., 2004) of incremental parsing.

In the first case, training set consists of pairs of words on all sentences RNC divided into two classes those which:

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 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:

- R – right list consisting of all unparsed words,
- L – left-hand list comprising the word for finding the relationship between R[0] and L[0],
- M – intermediate list. If relation between R[0] and L[0] has not been found, then the list M will be replenished with the word L[0].

Thus 4 classes of actions are used:

- No-Arc – transferring L[0] to the top of the list of M,
- Shift – moving the R[0], and all the words from M to L, so that the word of R[0] was the apex of L[0],
- Right-arc – setting a relation between R[0] and L[0],
- Left-arc – setting a relation between L[0] and R[0].

If the action is Right-arc or Left-arc, the word L[0] moves to M-list and becomes its apex. The training set in this case consists of examples corresponding to these four actions according to the type of sinto.

2.3 Determining syntactic relations

Methods MLP, SGD (Zhang T., 2004), SVM strategy one-vs-all (Rocha A., Goldenstein S., 2013), PNN, GNT (Sboev A., et al, 2012), ensembles of decision trees (RFC) (Breiman, 2001) in combination with the methods of reducing the dimension of the input space (Nystroem) (Kumar S., Mohri M., 2009) were investigated to determine the syntactic relations.

MLP neural network is trained by Error Back-Propagation algorithm. The number of neurons in the hidden layers is selected using a genetic algorithm. Neurons of hidden layers using transfer functions, such as sigmoidal $f_1(x) = 1/(1 + e^{-\alpha x})$, or tangential $f_2(x) = \tanh(x/\alpha)$, where α – slope parameter of activation function.

SGD is a method of selecting the parameters of the function (1) with (2)

$$f(x) = w^T x + b, \quad (1)$$

where x_i – input example, *w* – parameters of the model, *b* – coefficient.

$$E(w, b) = \frac{1}{n}\sum_{i=1}^{N} L(y_i, f(x_i)) + \alpha R(w), \quad (2)$$

where y_i – the desired output of *i*-th example, *L* – loss function, *R* – regularization term (measures l2 or l1), α – a positive coefficient (hyper parameters).

Various functions of two arguments *y* and $p = f(x)$ can be used as *L* (3-6):

1) $L(p,d\ ) = \max(0,1 - pd\ ) = \begin{cases} 1 - pd\ , d\ \leq 1 \\ 0, pd\ > 1 \end{cases}$, (3)

2) $L(p,d\ ) = \ln(1 + \exp[-pd\ ])\ pd\ )$, (4)

3) $L(p,d\ ) = \max(0,1 - pd)^2 = \begin{cases} 0,1 \leq pd \\ (1 - pd)^2, -1 \leq pd \leq 1 \\ -4pd \geq pd \end{cases}$, (5)

4) $L(p,d\ ) = |d - p|$, (6)

Classificator at the N classes is based on SVM (1) with One-vs-all strategy. It consists of N binary classifiers. They solve the problem of choosing one of all classes (Cl): $\mathrm{Cl_i}v(Cl_1, \ldots, Cl_{i-1}, Cl_{i+1}, \ldots, Cl_N)$. The class is selected by the maximum probability of determining by all the independent binary classifiers.

PNN is probability neural network.

$$G(x; x_i) = \exp\left[\frac{-1}{2\sigma_i^2}\sum_{k=1}^{P}(x_k - x_{ik})^2\right], (7)$$

x_i– vector of values of the i-th neuron of network, σ_i^2– dispersion, P – the size of the input vector.

Computational elements in PNN correspond to the values of input training examples. When testing the probability of the class of the input sample is determined using a Gaussian kernel (7).

GNT belongs to a class of neural networks with self-organizational process of learning that is based on winner-take-all-strategy. On the training phase the neuron that is the closest to the current input example wins and moves to the direction of the current object. On testing the input example gets the cluster name corresponding the name of the winner neuron.

We developed a complex model for determining syntactic relations that have small number of examples in training set. This model is based on the GNT and PNN. The basic idea of this model is to reduce the dimension of the training set on the training phase by calculating the center of mass of the cluster examples. It is performed only for the clusters that have examples that not relevant to the considered of Sinto. On testing phase examples parsed by PNN.

RFC is classification method based on an ensemble of decision trees. We explored different quantities of decision trees (10 to 1000). In all cases we uses Nystroem algorithm for reducing the dimension of input space of example for RFC.

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

1. Creating a sequence of classification neural network models to determine the 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.

2.4 Measures of accuracy of the syntactic parsing procedure

Further construction of the model of syntactic parser for the Russian language is performed on the basis on selected approaches to build the syntactic tree and to the formation of a training sample. Its accuracy is evaluated according to the following values:

- UAS – unlabeled attachment score;
- LAS – labeled attachment score;
- TRD –true root determination. It is the ratio of number sentences with correct root determination to number of sentences;
- TSSP –true structure of syntactic parse without type of syntactic relations. It is the ratio of number sentences with right structure of syntactic parse tree to the total number of sentences;
- TSPT –true syntactic parsing tree. It is the ratio of number sentences with right syntactic parsing tree to the total number of sentences.

3 Experiments

3.1 Estimation of accuracy of determining syntactic relations and the choice of method for constructing a tree of syntactic parsing

Enumerating all possible combinations of words in the sentence and the definition of syntactic relations in them

In this case, procedure for determining the syntactic relations included the development of neural network models for two tasks:

1. Filtering a set of examples that form syntactic relations (650 thousand) from those which do not form (10 million).The following models demonstrated the best results:
    - MLP (2 layers: 22 and 22 neurons), with estimations of PPV and NPV are equal 83.45/88.3%;
    - MLP (1 layer: 50 neurons) with estimations of PPV and NPV are equal 82.1/89.12%.
2. Determining of the syntactic relations for the examples within the filtered set. The following methods were investigated for this:
    - constructing separate neural network models to solve the tasks of binary classification for each type of syntactic relations;
    - forming groups of syntactic relations based on the accuracy of their definitions.

The models based on MLP (2 layers) and SVM using method of stochastic gradient descent (SGD) (Bottou, 2010) have demonstrated best results in solving the problem of independent determination of syntactic relations. Average PPV was 92.52 and NPV was 94.31%.

Average PPV in case of classification of examples of syntactic relations with small number of examples in training set with use of model based on GNT and PNN is equal 99.1% (NPV is 99.8%).

We achieved the best result when created 7 groups of types of syntactic relations. Last group consists of syntactic relations with small number of examples in training set. For their classification we also use the complex model (GNT and PNN). For other group best results demonstrated methods on base of SVM with training algorithm SGD. The use of grouping increases PPV to 95.6% (NPV to 97.4%).

Thus, the overall PPV of Sinto determination after selection syntactically significant variants is 79.89%. This estimate was obtained after processing the test sets by model of definition syntactically significant variants with use of MLP along with further classification of Sinto using methods SVM with SGD, GNT with PNN.

An approach based on incremental parsing scheme

The training set based on the Covington parse scheme is about 1.35 million examples, of which:

- 700 thousand relate to classes meaning the type of action without defining syntactic relation;
- 650 thousand relate to classes meaning the type of action with defining syntactic relation.

Several approaches have been analyzed (see Table. 4):

- model based on ensembles of decision trees or SVM with strategies such as One-vs-all (see Table 5, variants of descent – "multiclass");
- the decision on the basis of a binary classification action when a classifier is constructed for each action. The example for the class actions that have already built classifiers are excluded from the training set. (see Table 5, variants of descent – "binary").

Direct usage of ensemble classification methods such as RFC requires a lot of RAM, so we used Nystroem (Mu Li et al., 2010) algorithm for compressing the input data space.

Variant of descent	Using methods	PPV	NPV
multiclass	SVM (linear kernel)	90.1	91.2
multiclass	Nystroem + RFC	83.8	84.1
binary	SVC+SGD (1-4)	87.3	88.1
binary	Nystroem + RFC and SVC+SGD (1-4)	89.2	90.1

Table 4: PPV and NPV for defining actions in incremental parsing scheme

Thus, the approach based on the incremental parsing scheme with the classifier based on SVM with linear kernel was selected for the implementation of the model for parsing the Russian language.

3.2 Estimation of accuracy of the syntactic parsing procedure

Syntactic procedure model was implemented on the basis of the approach described in the previous chapter. Parameters of p_sinto were extracted using neural networks SVM, MLP, PNN. The results of testing the implemented model and comparison with other systems are presented in Table 5.

	Task description	UAS (%)	TRD (%)	LAS (%)	TSPT (%)	TSSP (%)
Our estimation	Using SVC for the classification of activities and the selected parameter set (without the word-forms)	85.81	82.23	79.33	14.05	29.47
	Using SVC for the classification of activities and the selected parameter set (with addition of the word-forms)	91.73	88.84	89.39	35.91	52.38
Estimation from literature sources	ETAP-3	94.3	---	92.3	29.7	37.4
	Incremental parsing scheme of Nivre-eager. Set of parameters based on word, part-of-speech, and morphological features	89.4	---	83.4	21.8	33.3

Table 5: Estimation of syntactic parsing and comparison with estimation from literature sources

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

Our approach to the formation of the plurality of parameters is shown. Assessment conducted on the basis of the generated set shows the level of possible ambiguity in determining the syntactic relations between words (98.9%) and the accuracy of parsing the sentence as a whole (79.9%). The selected set of parameters includes the extracted characteristics. Extraction is carried out on the basis of morphological characters of words in a sentence using the neural network classification algorithms. The best accuracy in this demonstrated algorithms based on network PNN, MLP, and SVM. Researches on the choice of the method of forming a set of training examples, and optionally of algorithm for syntactic parsing have shown the effectiveness of the incremental parsing scheme due to the lower number of non-syntactic significant examples in the training sample. Applying this approach to the formation of a sample of training examples gives greater accuracy of establishing the syntactic relations among two words. The paper shows a method based on the use of GNT and PNN algorithm demonstrates good accuracy of determination of the syntactic relations, having a small number of instances. However, a model based on support vector machines is selected because of better determination of all syntactic relations in the total range of their applications.

Thus, the results of presented research using the combination of neural networks SVM, PNN, MLP for extraction potential relations, SVM for defining syntactic relations with combinations of selected set of parameters, and word forms demonstrate precision of syntactic parse of sentence (TSPT) of 35.91%.

This is higher than from the literature sources (see (Kazennikov A., 2010), (Sharoff S., Nivre J., 2011)) but far lower than the evaluated achievable accuracy. The use of a model of syntactic parsing based on a system of sequential transitions leads to the accumulation of error resulting from inaccuracy of the definition of each syntactic relations (or actions). An increasing of quality of each classifier for determining the syntactic relations essentially improves parsing in whole. Because of this the development of improved models of classifier for determining syntactic relations on base of new topologies of neural networks, such as long short-term memory (LSTM) networks (Hochreiter S., Schmidhuber J., 1997) (Sutskever I. et al., 2014) and neural Turing machines (Graves A. et al., 2014) with use of proposed parameters is the prospect for our further work.

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).

References

Bengio Y. et al. (2003). A neural probabilistic language model. *The Journal of Machine Learning Research, 3*, 1137-1155.

Bottou, L. (2010). *Stochastic Gradient Descen*. Retrieved 03 2015, from http://leon.bottou.org/projects/sgd

Breiman, L. (2001). *Random forest*. Berkeley: University of California.

Chen D., Manning C.D. (2014). A fast and accurate dependency parser using neural networks. *Proc. EMNLP*, 740–750.

Collobert R. et al. (2011). Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research, 12*, .2493-2537.

Graves A. et al. (2014). *Neural Turing machines*. Retrieved from http://arxiv.org/ abs/1410.5401

Hochreiter S., Schmidhuber J. (1997). Long short-term memory. *Neural Comput. 9*, 1735–1780.

Iomdin L. et al. (2012). ETAP parser: state of the art. *The collection of materials of the International conference "Dialogue", v. 2*(№ 11 (18)), pp. 119-131.

Kazennikov A. (2010). A comparative analysis of machine learning. *Papers from the Annual International Conference "Dialogue"*(9 (16)), 157-163.

Kübler S. et al. (2009). *Dependency Parsing Synthesis Lectures on Human Language Technologies*.

Kumar S., Mohri M. (2009). *Ensemble Nystrom Method*. New York: Courant Institute of Mathematical Sciences.

LeCun Y. et al. (2010). *Convolutional Networks and Applications in Vision*. New York: Courant Institute of Mathematical Sciences. Computer Science Department.

LeCun Y., Bengio Y.,Hinton G. (2015). Deep learning. *Nature 521*, 436–444.

Mu Li et al. (2010). Making Large-Scale Nystr¨om Approximation Possible. *Proceedings of the 27 th International Conference on Machine Learning*. Haifa , Israel.

Nivre J. (2004). *Incrementality in Deterministic Dependency Parsing*. Vaxjo, Sweden: School of Mathematics and Systems Engineering.

R. Collobert. . (2011). AISTATS. *Deep Learning for Efficient Discriminative Parsing*.

Rocha A., Goldenstein S. (2013). Multiclass from Binary: Expanding One-vs-All, One-vs-One and ECOC-based Approaches. *IEEE Transactions on Neural Networks and Learning Systems*.

Russian National Corpus. (n.d.). Retrieved from http://www.ruscorpora.ru/instruction-syntax.html

Rybka R. et al. (2014). 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*.(2), 117-124.

Sboev A., et al. (2012). Neuronetwork package Neurotree adapted for the segment of the Russian grid network. *Informatization and Communication*.

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

Sutskever I. et al. (2014). Sequence to sequence learning with neural networks. *In Proc. Advances in Neural Information Processing Systems*, 3104–3112.

Wong Chun Kit. (2004). *Recursive Auto-Associative Memory as Connectionist Language Processing Model-Training Improvements via Hybrid Neural-Genetic Schemata*. Hong Kong: City University of Hong Kong.

Zhang T. (2004). ICML 2004. *Solving Large Scale Linear Prediction Problems Using Stochastic Gradient Descent Algorithms*.