Single-sentence Readability Prediction in Russian

Nikolay Karpov, Julia Baranova, Fedor Vitugin

National Research University Higher School of Economics,

Nizhny Novgorod, Russia

nkarpov@hse.ru,{ligros7,fedor.vityugin}@gmail.com

Abstract. In an effort to make reading more accessible, an automated readability formula can help students to retrieve appropriate material for their language level. This study attempts to discover and analyze a set of possible features that can be used for single-sentence readability prediction in Russian. We test the influence of syntactic features on predictability of structural complexity. The readability of sentences from SynTagRus corpus was marked up manually and used for evaluation.

Keywords: natural language processing, text readability prediction, single-sentence readability, syntactic links.

1 Introduction

One of integral parts of language teaching is reading, which gives some technical difficulties for professors and students. These difficulties are mainly connected with searching and understanding texts of a concrete level of difficulty (corresponding to a student's knowledge). At the moment there are several research projects which focus on the obtainment of text with the readability level needed for the education purposes. First approach is to classify texts with respect to its level and to retrieve the text needed. Second approach is to take any text and simplify it to the target readability.

This paper describes the part of a project which aim is to develop a system with a simplification functionality. It should be a system of text adaptation to a target level in Russian language as a foreign language (RFL). In the framework of project realization of the automatic simplification of texts in accordance with the language level, we were solving the identification problem of the source and the resulting levels of difficulty of the sentences or texts. Further step will be their lexical and syntactic simplification. In this study we give the results of application of a number of models that identify the level of difficulty of the text or single-sentences using different statistical parameters.

In section 2 **Related work** we presented an overview of the work of the researchers involved in the subject classification of the texts on the basis of complexity of reading in Russian, English and French. In section 3 **Text readability prediction** there are classical models and a model developed specially for the Russian language were adapted to our resources and was tested on texts of several levels of difficulty. Section 4 **Sentence classification** describes the results obtained by applying the formula of Flesch-Kincaid and Dale-Chall to identify the lexical and

structural complexity of the Russian sentences. In section 5, **Sentence classification with syntactic features** we give one of the variants of the model for the effective identification of readability of Russian sentences with the use of syntactic features. In section 6 **Conclusion** there are general conclusions on the executed experiments and plans for further work and improvement of the models considered.

2 Related work

The first studies on text complexity started in the 20's of the past century. This field of research was mainly developed in the field of work relating to the English language, but over the last decade a number of works related to other languages were worried out, that testifies to the fact, that the research relating to automatic identification of the complexity of the text is still highly relevant.

The complexity of reading can be represented as a function that finds a correspondence to a certain level of complexity of the predefined text with a variety of variables, extracted from the text. Traditionally variables allocated for the characterization of these texts are divided into two groups - lexical parameters and syntactic parameters. In one of the most common formula - the formula of Flesch-Kincaid [1], [2] the complexity of the text is represented as a linear function of the average number of syllables per word and the average length of the sentence.

A formula of Dale and Hall [3] also defines a syntax difficulty of the text as the average length of the sentence, but for lexical metric it uses the percentage of words not from list of 3000 Easy Words, which is based on familiarity of words. This means that all the words in the list are familiar to US children in the 4th grade.

With the growth of computing power there appeared an opportunity to build more complex models. Model of Collins-Thompson and Callan (2005) [4] uses frequency of words unigrams (a dictionary is specified for each level of the language) and features that some words are most possible for prediction a certain level of complexity of the text. Schwarm and Ostendorf [5] use more complex syntax parameters - the average height of trees parsing, the number of nominal and verbal groups, the average number of non terminal nodes and so on.

Automatic identification of a reading difficulty in Russian language is also researched in a number of works. Oborneva (2006) in her work [6] adapts the formula of Flesch and Flesch-Kincaid for the Russian language by means of adjustment coefficients: she compares the average length of syllables in English and Russian words and percentage of multi-syllable words in dictionaries for these languages. It is also worth noting the study of Krioni, Nickin and Philippova who define the complexity of educational texts in Russian language highlighting a number of more complex parameters of assessed texts: connectivity, structure, integrity, functional and semantic type, information, abstractness of the text presentation and complexity of linguistic structures [7].

Due to the large amount of research dedicated to readability assessment, we have highlighted only the most eminent works. Nevertheless all of them identify the difficulty of reading of the whole text. Our goal is to determine the efficiency of the developed techniques in relation to the texts in general and sentences in particular in Russian language as well as checking our own developed model to determine the difficulty of sentence reading.

3 Text readability prediction

First task was to perform the prototyping of Russian text retrieval with needed readability. The main goal of this process was to find which kind of variables and classification algorithm would allow us to obtain the highest indicators of precision and recall of readability prediction. There was conducted a series of experiments on the training of different classification algorithms. We experimented with the following algorithms:

- naive Bayes;
- k-nearest neighbors;
- classification tree;
- random forests;
- SVM.

Evaluation was performed with the help of cross validation on the test part of our collection. We extract features from a collection consists of 219 texts divided into four groups. Levels distribution is following: A1 (elementary – 52), A2 (basic) – 57, B1 (first) – 60, C2 (difficult) – 50 according to levels described in Common European Framework of Reference for Languages (CEFR) [8]. The first three groups include texts, created specially for second language learners of Russian, with respect to their level of language knowledge on the basis of news articles[1]. Fourth group (difficult) consists of original news for native readers. We extract 25 variables from texts proposed in the previous works:

1. Average number of words in the sentence of the text.
2. Average length of one word in a sentence.
3. Text length in letters.
4. Text length in words.
5. Average sentence length in syllables.
6. Average length of words in syllables.
7. Percentage of words with number of syllables more or equal to *N*. We define *N* as each value from 3 to 6.
8. Average sentence length in letters.
9. Average length of words in letters.

[1] http://texts.cie.ru

10. Percentage of words with number of letters more or equal to *N*. We define *N* as each value from 5 to 13.
11. The percentage of words in a sentence, not included in the active vocabulary of A1 level.
12. The percentage of words in a sentence, not included in the active vocabulary of A2 level.
13. The percentage of words in a sentence, not included in the active vocabulary of B1 level.
14. The occurrence in the sentence of concrete parts of speech.

We mark seventeen parts of speech in the texts according to the list of grams in the OpenCorpora [9]: noun (NOUN), full form of an adjective (ADJF), short form of an adjective (ADJS), comparative (COMP), personal form of the verb (VERB), infinitive form of the verb (INFN), full participle (PRTF), short participle (PRTS), gerund (GRND), numeral (NUMR), adverb (ADVB), noun-pronoun (NPRO), predicative (PRED), preposition (PREP), conjunction (CONJ), a particle (PRCL), interjection (INTJ). We were interested in occurrence of parts of speech as proposed by Francois, 2009 [10].

We did not use some variables described in paper [11] due to adaptation to our texts. We did not use variable connected with paragraph because our texts are very short. Texts do not have syntactic markup that is why the concept of a phrase was not used either.

First experiment was a binary classification of readability: A1 versus C2, A2 versus C2, B1 versus C2. With the help of Classification Tree, SVM and Logistic Regression algorithms the accuracy we got was really high, it was almost equal to 1.

Second experiment for texts classification of four levels got lower accuracy. An example of accuracy of text retrieval with B1 level of readability is shown in Table 1.

Table 1. Results of texts retrieval with B1 readability level.

Method	Classification accuracy	F-measure	Precision	Recall
SVM	0.8092	0.7965	0.8491	0.75
Classification Tree	0.9905	0.9916	1	0.9833
kNN	0.8131	0.7333	0.7333	0.7333
Random Forest	0.9818	0.9667	0.9667	0.9667
Naive Bayes	0.8726	0.7890	0.8776	0.7167

kNN is a *K* nearest neighborhood method. Results received during the second experiment are worse than the first experiment with only two levels. Due to the fact that results of the Classification Tree method reached 99%, we can say that the obtained results meet the needs.

To analyze the effect of each variable for the texts discrimination into 4 levels we ranked it by information gain ratio [12].

Table 2. Texts variables ranked by information gain ratio (top 10).

Variable name	Information gain ratio
The percentage of words in a sentence, are not included in the active vocabulary of A1 level	0.105141
The percentage of words in a sentence, are not included in the active vocabulary of A2 level	0.105141
The percentage of words in a sentence, are not included in the active vocabulary of B1 level	0.084211
Percentage of words with 8 letters or more	0.040098
Percentage of words with 9 letters or more	0.038431
Percentage of words with 7 letters or more	0.036923
Average sentence length in syllables	0.034359
The average length of one word in a text	0.034359
Percentage of words with 10 letters or more	0.033689
Percentage of words with 5 syllable and more	0.033193

The highest information gain ratio have first three variables which are lexical ones. We can say that they are most important variables for discrimination.

4 Sentence classification

Next task was to make a prototype of an algorithm to retrieve difficult sentences for further simplification. This algorithm is based on a sentence classification with respect to its readability. For results evaluation we use subcorpus of Russian national Corpus (RNC) - corpus SunTagRus [13] which has morphological and syntactic metadata. We manually tagged 3500 sentences from this subcorpus to mark their structural level of perception complexity. We found out that level B1 suits the majority of our students. So, we created a binary sentence markup, which is 1) B1 or lower than B1 and 2) Higher than B1.

Lexical difficulty markup was made on the basis of active vocabulary of three levels: A1, A2 and B1. The most complete vocabulary list (B1) includes 2500 words. So, we defined sentences having more than 33% active vocabulary words as lexically difficult ones.

Thus, we have two kinds of markup: structural complexity and lexical difficulty. As an intersection of its lexical and structural level of difficulty we obtained markup of a total level of difficulty.

Dale-Chall model was developed to define the difficulty of text with the help of linear function of flowing variables: average sentence length (number of words divided by number of sentences) and rear words in the text.

When we use these variables for single sentence readability prediction we need to adapt them as following: sentence length rather than average sentence length, percentage of words not in the active vocabulary with respect to sentence length (number of words in the sentence not in the vocabulary divided by total number of words in the sentence) instead of rear words percentage. In our case, we don't need to use dictionary of the Russian language with the frequency of words occurrence because we have a definite list of words which are contained in active vocabulary.

These two variables were automatically extracted for each sentences in our corpus. We predict readability for single sentence using different methods of machine learning as shown in Section 3. Evaluation was performed with the help of cross validation on the test part of corpus.

To evaluate the influence of each variables first we try to predict difficulty using variables separately, next we predict it using both variables. It is easy to see that even in the case of prediction with the help of sentence length we can obtain good results. But if we need to classify to more than two numbers of levels, accuracy will decrease. Precision of difficult sentence retrieval is lower than simple sentence retrieval.

Accuracy of readability prediction on the basis of both variables is much higher. The second variable - percentage of words not in the active vocabulary cut off many difficult short sentences. It is effect to the precision of difficult sentence retrieval. The results are presented in Table 3.

Table 3. Results of readability prediction using variables: sentence length and percentage of words not in the active vocabulary.

Method	Classification accuracy	F-measure (difficult /simple)	Precision	Recall
Naive Bayes	0.8846	0.9242/ 0.7581	0.9378/ 0.7246	0.9110/ 0.7950
Logistic regression	0.8745	0.9212/ 0.6921	0.8945/ 0.7833	0.9495/ 0.6199
kNN	0.8941	0.9299/ 0.7840	0.9519/ 0.7318	0.9089/ 0.8441
Random Forest	0.8840	0.9208/ 0.7837	0.9747/ 0.6808	0.8725/ 0.9233
Classification Tree	0.8955	0.9308/ 0.7866	0.9527/ 0.7347	0.9099/ 0.8465

We have opposite situation in this case. Precision of difficult sentence retrieval is higher than simple sentence retrieval. We can conclude that even using only these two variables we can effectively predict sentence readability.

Flesch-Kincaid model grade level formula was also used to determine readability. The formula utilized in the software is [(0.39×ASL)+(11.8×ASW)−15.59], where ASL is the average sentence length (number of words divided by number of

sentences) and ASW is the average syllables per word (number of syllables divided by number of words). To apply this formula to the problem of estimating the difficulty of the single sentence we can save ASW in its original form and instead of ASL use sentence length (number of words).

If we come to analyze how the lexical difficulty is predicted with the help of the average syllables per word (ASW) it is easy to notice that ASW exert to classification accuracy of difficult sentences and not exert to simple one. The reason is that Russian language is characterised by the presence of many long words (with many syllables) which are simple ones because they are created by combining short words. This is the help of two variables (ASL and ASW) we get results which are shown in Table 4.

Table 4. Results of difficult/simple sentence retrieval from text using ASL and ASW.

Method	Classification accuracy	F-measure (difficult /simple)	Precision (difficult /simple)	Recall (difficult /simple)
Naive Bayes	0.7967	0.8794/ 0.3550	0.8119/ 0.6386	0.9590/ 0.2458
Logistic regression	0.7945	0.8770/ 0.3761	0.8156/ 0.6086	0.9484/ 0.2722
kNN	0.7746	0.8640/ 0.3434	0.8093/ 0.5094	0.9265/ 0.2590
Random Forest	0.7910	0.8788/ 0.2431	0.7961/ 0.6910	0.9806/ 0.1475
Classification Tree	0.7801	0.8669/ 0.3673	0.8140/ 0.5318	0.9272/ 0.2806

Total accuracy for only two variables is relatively high but the recall of simple sentences retrieval is quite low. Active vocabulary in the first certified level of Russian language could not be exactly determined using the average syllables per word.

5 Sentence classification using syntactic structure

We use deeper sentence features which potentially can improve accuracy of readability prediction - syntactic relations of words. Our experiment was carried out on the basis of SynTagRus corpus which has morphological and syntactic metadata. We decide to use syntactical features of a sentence as a basis of classification algorithm because this approach shows better results on the preliminary stage whether morphology features or n-gramms. In this case on the basis of syntactical features classification tasks look as follows. The sentences are tagged with morphological metadata using OpenCorpora [9]. On the basis of morphological marks we generate syntactical links. Its syntactical links help us to predict single sentence readability.

SynTagRus includes about 60 types of syntactic links grouped as it proposed in RNC. We try to predict sentence readability with the help of two data representation. First we use all 60 types of syntactic links. We get following experimental results shown in Table 5.

Table 5. Classification using 60 types of links.

Method	Classification accuracy	F-measure	Precision	Recall
Naive Bayes	0.7570	0.7459	0.7813	0.7136
Logistic regression	0.7112	0.7077	0.7160	0.6995
kNN	0.7286	0.7146	0.7531	0.6798
Random Forest	0.7582	0.7472	0.7822	0.7153
Classification Tree	0.7047	0.6414	0.8158	0.5284

Then we use aggregated links to 4 groups as it proposed in RNC. Classification accuracy using aggregated variables was lower. On the basis of obtained experimental results it was concluded that we should use all types of links without aggregation. The best precision and recall showed SVM algorithm.

It is obvious to assume that syntactic variables can predict structural difficulties better. Thus we used the same approach as it was with other previous models, perform experiment with structural and lexical difficulty separately. Results are presented in Table 6.

Table 6. Results of structural difficulties prediction using only syntactic variables.

Method	Classification accuracy	F-measure (difficult /simple)	Precision	Recall
Naive Bayes	0.8085	0.8021/ 0.8144	0.8244/ 0.7942	0.7810/ 0.8356
kNN	0.7681	0.7128/ 0.8055	0.9271/ 0.6965	0.5790/ 0.9550
Classification Tree	0.8180	0.8056/ 0.8289	0.8589/ 0.7860	0.7585/ 0.8768
SVM	0.7956	0.8010/ 0.7900	0.8972/ 0.8173	0.9174/ 0.7645
Random Forest	0.8374	0.8307/ 0.8436	0.8610/ 0.8170	0.8271/ 0.8719

We can conclude that syntactic variables allow to predict structural difficulties more efficiently than simple variables. Next we use all kind of variables (syntactic

and lexical) to predict total difficulty of sentence. As a lexical variable we use percentage of words not from active vocabulary of the corresponding level (Table 7).

Table 7. Results of total readability prediction using all kinds of variables and syntactic links.

Method	Classification accuracy	F-measure (difficult /simple)	Precision	Recall
Naive Bayes	0.8191	0.8906/ 0.4767	0.8354/ 0.6975	0.9537/ 0.3621
kNN	0.8224	0.8893/ 0.5501	0.8571/ 0.6493	0.9241/ 0.4772
Random Forest	0.9443	0.9640/ 0.8768	0.9620/ 0.8832	0.9661/ 0.8705
Classification Tree	0.9364	0.9584/ 0.8648	0.9679/ 0.8380	0.9491/ 0.8933
SVM	0.8633	0.9125/ 0.6875	0.9679/ 0.7165	0.9491/ 0.6607

Last approach gives more stable results and may be used to increase the number of classes of sentence complexity.

Table 8. Results of total readability prediction using all kinds of variables and syntactic links.

Variable name	Information gain ratio
The percentage of words in a sentence, are not included in the active vocabulary of B1 level	0.318
Sentence length in letters	0.122
Percentage of words with 3 syllable and more	0.119
Sentence length in syllables	0.118
Sentence length in words	0.098
Syntactic predicative link	0.095
Average words length in syllables	0.092
The average length of one word in a text	0.092
Percentage of words with 7 letters or more	0.069
Percentage of words with 5 letters or more	0.069

6 Conclusion

Classical models and models developed specially for Russian language were adapted to news texts retrieval. These models give good results. We managed to develop a precise classification system of news texts in Russian with respect to their readability.

Accuracy of four levels classification was lower. Due to the fact that obtained results of the Classification Tree and Random Forest methods reached 99-98%, we can say that they met our needs.

We adapted traditional classification techniques with statistical features like Flesch-Kincaid and Dale-Chall to identify lexical and structural complexity of Russian sentences. These techniques were tested on set of sentences where readability was manually marked as binary classification.

Finally, we found one of the variants of the model for the effective identification of readability of Russian sentences with the use of syntactic links. We found that syntactic features can predict structural complexity. Total set of features with statistical, lexical and syntactical ones can predict sentence readability with 0.9661 amount of recall using Random Forest algorithm. Most important features for this classification are lexical ones.

Acknowledgment

This study comprises research findings from the «Adaptation of texts from the Russian National Corpus for the electronic textbook «Russian language as a foreign one» carried out within The National Research University Higher School of Economics' Academic Fund Program in 2013, grant No 13-05-0031.

References

1. Flesch, R.: A new readability yardstick. J. Appl. Psychol. 32, 221 (1948).
2. Kincaid, J.P., Fishburne Jr, R.P., Rogers, R.L., Chissom, B.S.: Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. DTIC Document (1975).
3. Chall, J.S.: Readability revisited: The new Dale-Chall readability formula. Brookline Books Cambridge, MA (1995).
4. Collins-Thompson, K., Callan, J.: Predicting reading difficulty with statistical language models. J. Am. Soc. Inf. Sci. Technol. 56, 1448–1462 (2005).
5. Schwarm, S.E., Ostendorf, M.: Reading level assessment using support vector machines and statistical language models. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. pp. 523–530. Association for Computational Linguistics (2005).

6. Oborneva, I.: Automatic assessment of the complexity of educational texts on the basis of statistical parameters, (2006).
7. Krioni, N., Nikin, A., Filippova, A.: Automated system for analysis of the complexity of educational texts. Manag. Soc. Econ. Syst. 11, 101–107 (2008).
8. Verhelst, N., Van Avermaet, P., Takala, S., Figueras, N., North, B.: Common European Framework of Reference for Languages: learning, teaching, assessment. Cambridge University Press (2009).
9. Victor Bocharov, Maria Stepanova, Natalia Ostapuk, Svetlana Bichineva, Dmitry Granovsky: Quality assurance tools in the OpenCorpora project. Computational Linguistics and Intelligent Technology: Proceeding of the International Conference «Dialog–2011». pp. 10–17 (2011).
10. Francois, T.L.: Combining a statistical language model with logistic regression to predict the lexical and syntactic difficulty of texts for FFL. Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop. pp. 19–27. Association for Computational Linguistics (2009).
11. Nevdah, M.: Development of a method of automated evaluation of the complexity of educational texts for higher school. (2008).
12. Kent, J.T.: Information gain and a general measure of correlation. Biometrika. 70, 163–173 (1983).
13. Nivre, J., Boguslavsky, I.M., Iomdin, L.L.: Parsing the SynTagRus treebank of Russian. Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1. pp. 641–648. Association for Computational Linguistics (2008).