Automatic identification of lexical and syntactic complexity of the data on the basis of National Corpus of the Russian Language

Yulia Baranova
Higher School of Economics
ligros7@gmail.com

Tatyana Elipasheva
Higher School of Economics
knopi4ka@gmail.com

Abstract

In this paper will be discussed methods of automatic classification of sentences in Russian by category of lexical and syntactic complexity followed simplification and applying the information about their structure to search for additional materials from open electronic resources to broaden the base of expansion adapted examples for textbook of Russian as a foreign language.

1 Introduction

The basic idea of the research is associated with the ability to automatic simplify or adapt texts from the Corpus of the Russian Language and open digital sources in accordance with the level of language proficiency. The task can be defined as the automatic identification of lexical and syntactic complexity of the data and automatic conversion of complex proposals to simple to develop base adapted examples for a textbook of Russian as a foreign language.

2 Research matter and actuality

The actuality of this project is that the base of the textbook is a material of National Corpus of Russian with extraction of the most frequency language constructions. Nowadays two types of electronic textbook exist: scanned traditional textbook and textbook with the hypertext. The main advantages of such kind of textbooks are limited by opportunity to send them by means of e-mail and to keep them on the flash and hard drives. The amount of interactive electronic textbook is few and most of them have unfriendly interface.

The textbook is not only a collection of words and rules it must learn how to use the language in communication and in academic researches. The material of National Corpus of Russian makes the process of language learning more communicatively oriented, helps to note the most typically occurring modifications that affect different functional styles of language.

The main goal of students that want to learn Russian is to obtain professional knowledge using the foreign language. For that reason the teaching of reading and writing is increasingly important on the basis of material that consists of general scientific and formal texts. Therefore new technologies must be implemented for the teaching of Russian as a foreign language, new approaches and methods must be developed that will allow saving Russian education on the high level. The total volume of National Corpus of Russian is more than 340 million of word usages; it provides the necessary variations of material for composition of exercise. This material consists of not only imaginative literature but formal, scientific, publicity texts too.

3 Determination of lexical levels

The second part of this project is to define the lexical levels and to compose the vocabulary in order to determinate the sentences for corpus. In scope of part the following purposes were established:

- To organize the group of experts in Russian syntax for the determination of lexical levels L1 and L2 and the development of methodology for composition of electronic textbook of Russian as a foreign language.
- To investigate the main algorithms of text adaptation in terms of linguistic rules
- To have a practice with syntactic parsers for Russian, concordancers, frequency dictionaries
- To advance their features for adaptation and simplification of language specific structures

First of all, the group of experts composed the dictionary, i.e. lexical minimum, on the basis of

current textbooks used in education. Two levels were defined: the basic level that contains 945 words and the first level that contains 840 words. All words are represented in the initial forms and some conjugates can be found in the lexical minimum.

The main purpose of lexical minimum is to define which sentences can be included to corpus for composition of practical exercises in future. After the automatic classification of all sentences in corpus (the first step) one of the output sets represents simple sentences but with lexicon of different complexity levels. In order to include the sentence in corpus it passed a lexical test. During this test the proportion of words from lexical minimum against total amount is calculated in percentage terms by the script and if it is equal to 70% or more then this sentence will be included to corpus. For this stage of project the following experiment was executed.

The corpus SynTagRus was used as a test set where all texts are morphologically and syntactically marked up. The syntactic structure of sentence has a representation in the form of dependency tree where the set of nodes is words from the sentences and branches describe the relationships between nodes. SynTagRus consists of structures that are disambiguated morphologically and syntactically. This means that each word from the text has only one morphological structure and each sentence from the corpus has only one syntactic structure at the same time. SynTagRus consists of texts of three main types: modern Russian prose, popular science and public policy articles, and texts from news bulletins. The group of experts developed the rules based on the morphology and syntax according to which one or another sentence can be marked up as simple or complex. The next stage of experiment was to divide simple sentences from complex by script keeping the SynTagRus data structure in safe. The output corpus of simple sentences was submitted to lexical minimum test. As a result of experiment the volume of the output corpus is not representative.

4 Determination of the structural complexity of sentences

The second stage of processing the material is to analyze the structure of the sentence to determine whether it level of complexity. At this stage of the development of all the sentences are divided into two categories - simple and complex. This division is carried out in accordance with the definitions of these categories in the syntax of the Russian language. As the input unit is used not marked up sentence which should assigned to one of the specific categories of structural complexity by the application being developed. To solve the problem of determining the complexity of the proposals will take either of two approaches: machine learning based on the training set and pre-defined expert rules based on the morphology of Russian language.

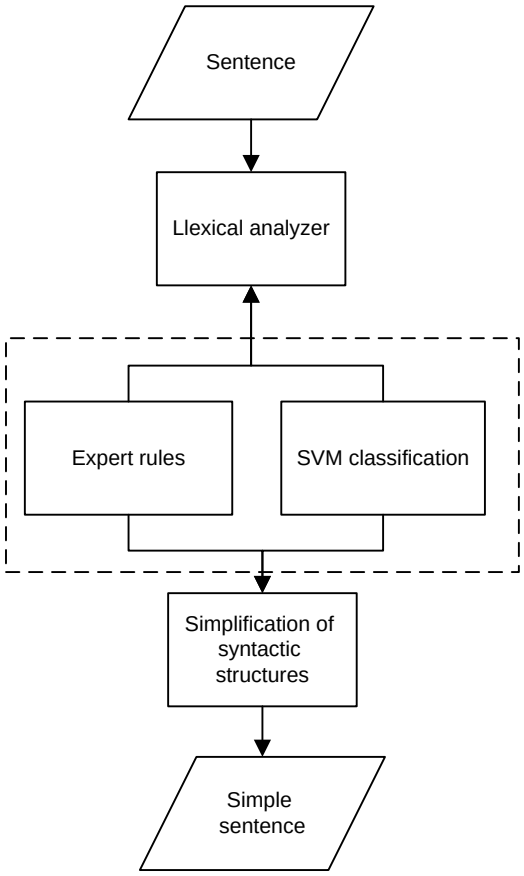

Figure 1: Block diagram of the simplification programming module

In the future, based on the results of evaluation of the recall and accuracy of the results it will be decided which of these methods should be used, or it will be a combination of both methods into a single evaluation model with certain weights.

4.1 Using of the expert rules

Expert rules are drawn up on the basis of the syntax and morphology of Russian language. These rules are defined a set of complicating structural blocks for sentences. Later in the evaluation of suggestions these rules are used as templates. Also in further using these patterns will simplify

sentences by removing complicating structures of them.

Each rule consists of a sequence of parts of speech, punctuation, and signs position in the sentence. They are all programmed using a language with module for morphological parsing. To use the expert rules it is necessary to implement preprocessing of the text. All of the words in the sentence should be marked by appropriate tags of morphological markup. This markup and punctuation are converted to the structure of the sentence, which is then compared with the already defined templates of expert rules. In case the sentence is equal to one or more structural template it is considered to be complicated.

As a result of the algorithm each sentence refers to one of the categories. To verify the completeness and accuracy of expert rules it will used a test sample in which each sentence is assigned manually to a category.

4.2 Training of the classifier

The classifier will be constructed on the basis of methods of support vector machine, which are implemented in the NLTK (Natural Language Toolkit) library for Python 2.7. The main idea of support vector machine methods consists in treating each feature as a dimension and positions features in N-dimensional feature space. The next step is determining optimal hyper line which separating features with the maximal gap in this space. According to the results of the division of space hyper lane determines class to which object is belong. Currently in NLTK this algorithm is implemented in binary form only, but it is enough for solving our problem, because at this stage of development we need to learn how to separate sentences only by two classes of complexity.

As the training set we used morphologically marked up part of National Corpus of the Russian Language. Total volume of the corpora is about 4000 sentences. Training is carried out by 80% of the corpora; the rest will be used to evaluate the effectiveness of training.

The unit of processing is a sentence that has the tag "simple" or "complex". These sentences have been annotating each manually at least by two experts. In case of discrepancy between expert opinions about the syntactic complexity of the sentence, decision shall be made by a third party. This provides the minimization of errors of manual marking of the corpora.

In addition to the tag level of sentence has been entered the tags of the each word`s level. This entire structure of marking will be stored in the form of XML files; it will store the tags for the words and sentences in the attributes of an appropriate level. Morphological marking of words is carried out automatically using the open-source library pymorphy2 for Python programming language, the interaction with XML files is carried out using some standard tools of this language.

After training on a sample quality of training is being tested. It uses 20% of the corpora that are not used before in training. On the basis of these results it is concluded feasibility of using machine learning techniques to recognize the syntactic complexity of sentences in Russian.

4.3 Using of the resulting structures for extracting an additional examples of sentences

The initial results of experiments for searching sentences which satisfy both lexical minimum and syntactic complexity demonstrate that they are less than 10% of the corpora volume.

In order to increase the sample size as well as to test the classification algorithms and validation algorithm of entering into the lexical minimum, we can use search sentences according to patterns. In this case, the template is morphological structures of manually marked earlier sentence that the experts identified as simple. According to these structures we can search new simple sentences in the public domain of the Internet. For this search the following algorithm can be used in:

1. As a set of patterns is accepted morphological structures of sentences that are marked by experts as simple. Structures are stored in the format used by the morphological analyzer (pymorphy2).
2. Specifies the set of sources on which the search and retrieval of text data is perform.
3. The extracted data is split to the sentences, each of which is processed by the morphological analyzer.
4. Each of the new obtained structures is compared for a matching with the database structural patterns. We consider only total matches.
5. New sentences which match with the patterns will be checked by experts.

This will reveal the markup errors and patterns, which morphological structure does not allow determining the complexity of the proposal exactly.

5 Simplifying of syntactic structure of the sentences

At the stage of simplifying the syntactic structure we need to remove the complicating syntactic structures from these sentences. Of course, this should be done only for sentences, which syntactic structure is determined automatically as complicated. Due to the basis of expert rules, we can recognize the complicating structures by the morphological structure of the sentence and punctuation and remove them.

At this stage, remove all complicating syntactic structures it is no matter how obtained simplified sentences equals to the original the completeness of the transferred sense. Subsequently, for this purpose should been introduced an additional processing stage, based on semantic analysis of the sentence.

In parallel with expert rules will be verified the effectiveness of methods of simplification sentences through syntactic parser. Corpora used for this work also include syntactic markup that can be used for machine learning. With the help of a trained syntactic parser, we can build syntactic tree received any sentences. Thus, the task of simplifying the sentence submitted as a parse syntactic tree would be in cutting branches at a certain level.

Conclusion

The project is under development now and not yet completed. At the moment, the following works have been completed:

1. Compiled dictionary of lexical minimum of Russian of the first and second levels. Currently is used only lexical minimum of the first level.
2. Algorithm for analysis of the sentences for compliance with the lexical minimum is described.
3. SynTagRus is analyzed for a matching to the lexical minimum, the results show that the number of suitable examples for training does not exceed 10% of the corpora.
4. SynTagRus is marked up manually into simple and complex by the syntactic structure of the sentence, base of expert rules with the structures of complicating elements is composed.
5. On the basis of the markup by complexity we will investigate the quality of the algorithms for determining syntactic complexity of the sentences. Currently will be investigated two methods: a support vector machine (machine learning) and using an expert rules.
6. An algorithm for extracting further examples of necessary language level sentences from open web-page on the Internet by defined patterns of simple sentences has been developed.
7. Analyzed the possible algorithms for automatic simplification of syntactic structure of the sentence using syntax trees and expert rules.

It is also planned to develop a web interface for the application, which will provide the ability to open access to all obtained learning resources, including electronic textbook of Russian as a foreign language.

References

Dobrushina Nina. 2009. *Corpora methods in education Russian //* Russian National Corpus. SPb (in Russian).

Steven Bird, Ewan Klein, Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media, Inc.

Boguslavsky I.M., Grigorieva S.A., Grigoriev N.V, Kreidlin L.G, Frid N.E. (2000). *Dependency Trebank for Russian: Concepts, Tools, Types of Information.* //Proceedings of the 18th Conference on Computational Linguistics. Vol 2, 987-991, Saarbrücken.

Chardin I.S. (2001). *Using a tagged corpus to resolve syntactic ambiguity in the ETAP-3 Linguistic Processor*. // Proceedings of the 2nd All-Russian Conference "Theory and Practice of Speech Resources"(ARSO-2001). Moscow State University, Moscow. [in Russian]

Sichinava, D.V. (2001). *On the problem of building Russian linguistic* corpora *for the Internet*. URL:www.mccme.ru/ling/mitrius/article.html [in Russian]