Hybrid Algorithm of Classifying Candidates for Subject Area Terms

I.A. Andreev, V.S. Moshkin, N.G. Yarushkina

Extracting terminology from text

A significant disadvantage of text processing by the Mystem program is the definition of more than one variant of word usage in one position. This drawback is a significant obstacle to obtaining accurate data when decomposing sentences into phrases. To eliminate this shortcoming, rules for disambiguation (by parts of speech) were developed Some of the rules are presented in Table I.

Table I. Disminumination rules (example)

Part of speech 1	Part of speech 2	Result
Noun	Verb	Noun
Noun	Adjective	Adjective
Adverb	Noun	Noun
Gerund	Noun	Noun
Pronoun	Noun	Pronoun
Numeral	Noun	Numeral
Pronoun	Conjunction	Conjunction

For the possibility of extracting terms from the texts of the subject area, linguistic templates were developed, with the help of which it is possible to select the main terms. In Russian, the syntactic structure of domain terms in more than 90% of cases corresponds to the five patterns shown in figure I.

Figure I. Patterns of terms

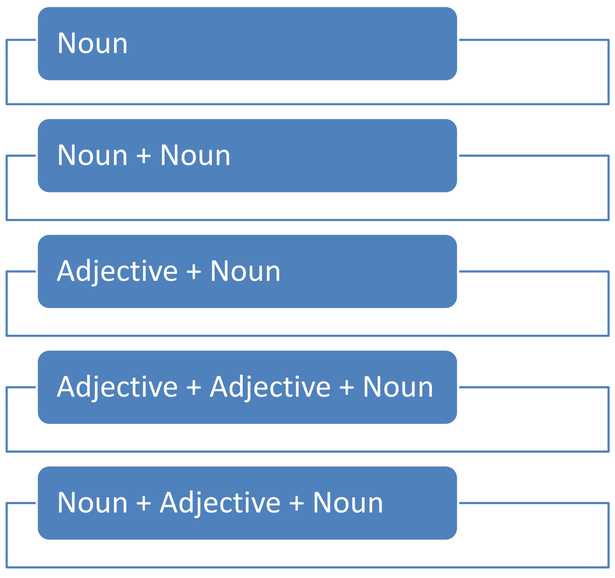
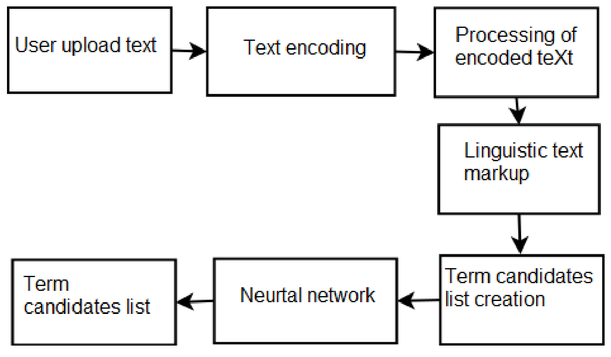


Figure II. The algorithm of the system



Experiments

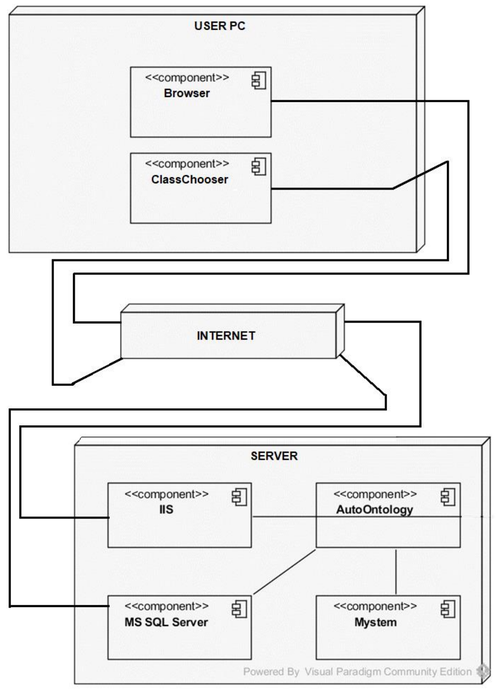
As input data for the experiment, a text of 40,000 words on the topic "Time Series" was chosen. As a result of the system operation, a list of terms was obtained, which allows an expert to study the results in more detail. Before work, the neural network was trained on a test set.

Table II. Characteristics list of terms

Term List Source	Term count	Error count	Undiscovered terms
Expert	2157	0	52
Software	2286	97	20

Based on the data obtained, the quality of the software was calculated. The percentage of error-free definitions of terms was 100%. This is due to the fact that, despite a number of errors and vague terms, the software determined the terms that the expert missed when subtracting.

Figure III. Scheme of the software used in the experiments



Conclusion

This paper aims to develop an algorithm for automated generation of a term list related to a subject area and assess the quality of the generated list. The automated compilation process and the developed algorithm are described. The development of the subject-based term list consists of several stages.. Upon completing each stage intermediate data is generated to be used in the future. The method is based on linguostatistical data obtained from the text and the "black box" of the neural network. To assess the quality of the generated list, a team of invited experts has compiled term lists for the texts used in the experiment. The comparison shows that the information system provides results rank close to those created by the experts. However, due to a number of errors made by the system, the results are inferior.

Future research consists of new linguostatistical method integration and adjustment of neural network's parameters.

ITNT-2022
151, Molodogvardeyskaya st., Samara, Russia (itnt-conf.org)

Ulyanovsk State Technical University

Ulyanovsk State Technical University
32, Severny Venets st., Ulyanovsk, Russia (www.ulstu.ru)