ABSTRACT

Title: Algorithm elaboration of terminological information extraction from the scientific and technical texts

Author: Mukhamed N. Gaunov

Research supervisor: Associate Professor, Ph.D. Nikolai A. Knyazev

Research initiator: State budgetary institution of additional professional education "Center for Continuous Development"

Topical importance: any linguistic task will cause a large number of problems when using software to solve it. Nowadays, statistical algorithms are used to a greater extent. Graduation qualification thesis suggests combining the statistical method and linguistic post-processing to reduce the probability of error when searching terminological information in thematic texts.

Goal: the identification of specific features of terms as lexical units within the text for the correct operation of the linguistic part, and the subsequent creation of an algorithm for searching terminological information in scientific and technical texts based on them, and then combining this algorithm with the statistical method in software.

Tasks: the study of terminological units of a certain type for the presence of specific characteristics. Creation of an algorithm based on the identified characteristics. Search for a suitable statistical algorithm. Creating a neural network. Adapt the network to the search, its connection to the algorithm. Training of the neural network for solving the task.

Theoretical value: the theoretical significance of the work is to justify the need of synthesizing linguistic and statistical methods.

Practical applicability: practical significance of the work done is to use the created software for compiling glossaries, terminological dictionaries, and also for use as part of a larger data mining analysis.

Results: A large amount of information and material was studied, the goal was achieved and the tasks were accomplished. An intelligent system of automated replenishment of the database of terminological units was developed. Also, software for this system was implemented using a high-level programming language - C #. As the extracted information, relative linguistic units were chosen - terms that can help in solving most linguistic tasks. The final version of the program showed that in 80% of cases the program correctly defines the language units. There is a definition of extra words, which is caused by a small number of analyzed and saved texts.

Implementation advice: The program completely copes with the task. At this stage, the product adds to the database some words that are not terms. For maximum efficiency of the program it is necessary to work with a large number of texts, which will affect the IDF parameter, which directly affects the final result for the better. Also in the future, more thorough and profound work is planned on the dictionary of tokens. It is necessary to identify the shortcomings that arise when defining interdependent and mutually exclusive tokens in order to coordinate the performance of the neural network much more efficiently.