The automatic text analysis toolset for solving practical and research tasks

Ekaterina Politsyna¹, Sergey Politsyn²
Moscow Aviation Institute (National Research University), Volokolamskoe sh., bld. 4, Moscow, Russia, 125993
E-mail: ¹kathrin.beaver@mail.ru, ²pul_forever@mail.ru

Abstract. Tasks of automated text analysis are increasingly moving from theoretical research and experimental systems into practice. Increased volumes of text information for manual processing caused the need to use automated tools to solve a variety of tasks for a wide range of users operating not only desktop computers but more and more often mobile devices. There is the need of creating new automated systems and services for a wide variety of applications and the lack of universal algorithms for text analysis. In addition, in connection with different performance requirements of the large data volumes processing, it is urgent to develop integrated tools of automated text analysis, quickly solve user problems on any device, tools for deeper text study and create a basis for further research in the field of computational linguistics and testing of new algorithms of analysis. The proposed toolset includes an open system of automated text analysis, web-portal "Automated text analysis," a set of services and mobile applications. Its use will provide an ability to create and use software tools to solve practical problems, taking into account the interests of a wide range of users, and promotes research in the field of computer linguistics and deploying their results into various applications.

1. Introduction

The automated text analysis tasks have passed from the field of theoretical research and experimental systems into practice. For several decades, many scientists have proposed various approaches for solving practical problems of automated text translation from one natural language to another, classification and clustering, search, etc., also many programs which implement them have been created [1–5].

They are based on the results of many works in the field of computer linguistics: Y. D. Apresyan, M. Minsky, I. A. Melchuk, D. A. Pospelov, R. Shenk, I. Wilks, G. G. Belonogov, A. K. Zhokhovsky, C. Fillmore, D. Jurafsky, M. G. Malkovsky, V. A. Fomichev, N. N. Leontiev, A. V. Sokirko, A. A. Kretov and many others. These works are devoted to the creation approaches of formal representation of natural language and their automatic construction, algorithms for morphological and syntactic analysis of the text, algorithms of extracting various statistical Characteristics of texts, analysis of large amounts of text data, etc.

The increased volumes of text information processed by people leads to the need of automated tools to solve a variety of problems for a wide range of users, using both computers and mobile devices. People exchange messages on social networks and instant messengers, read literature, news, reviews and other information, write books and articles, prepare reports on work results, look for information on reference and encyclopedic resources.
The results of automated text analysis are demanded not only for scientific problems and processing a large amount of data, but also for everyday activities for professional and personal purposes: not only for searching and translating into other languages, but also for SEO-optimization, quick topic recognition and text summarization, spelling and punctuation, summarizing the text to adapt it to the appropriate audience, speed up perception or display on a mobile device, etc.

Despite the variety of approaches to solving problems of automated text analysis, there is no universal solution to each of them for their implementation in software systems and applications for a wide range of users. In most cases, these are either experimental systems or separate services created to solve a small range of problems. There are architectures and sets of libraries and tools for creating, researching and using a wide range of different analysis models, as well as integrating them with information retrieval and storage technologies. For the English language, systems have been created that allow combining the sets of provided text analysis tools, but they are implemented as a set of libraries for software development and practically do not support working with texts in the Russian language (GATE, UIMA, Apertium, LingPipe) [6–8].

The need to create automatic systems and services for solving a variety of problems, the lack of universal text analysis algorithms, and the need to use computer linguistics in applied systems make it urgent to develop a set of automated text analysis tools with a set of services for solving problems by any user rapidly, tools for deeper text research and testing of new algorithms and a set of software tools for use in the industrial software.

2. The Structure of the Toolset
The set of tools is a platform for research in the field of computer linguistics, a set of services for solving individual tasks and a framework for developers of new text analysis tools for applied software systems. The structure of the toolset is shown in figure 1.

![Figure 1. The structure of the toolset.](image-url)
The set of tools is based on the open automated text processing system for the Russian language, which includes a number of text processing tools, accumulation tools for the information received and its subsequent analysis [9]. The system includes a client-server multi-user web application with graphical user interface for accessing basic and analytic processing tools and a set of methods of the program interface for implementing services that solve the most common computer linguistics tasks. The open automated text processing system is designed to conduct research in the field of computer linguistics, develop custom text processing algorithms without the need to write code, test new algorithms to improve the core of the system, etc.

The special management system has been developed to manage data structures extracted from texts, store and edit analysis scripts, and administrate user file space.

The TAWT framework was developed [10] based on the algorithms implemented in the core of the system and optimized for highly loaded industrial systems, this framework contains a set of tools that implement the main stages of automatic text analysis for the Russian language – graphematic, morphological, semantic-syntactic, united by common data structures and having common APIs, which provides the opportunity to start working on the Quick-Start principle and the ability to replace analysis components if necessary.

On the basis of the system and the framework, a number of services have been created for solving classification tasks, extracting keywords with different methods, obtaining statistical characteristics of texts, and summarization. All services are available on the Automated Text Analysis portal (textanalysis.ru) and provide not only a graphical user interface, but also REST API with a set of corresponding methods. Using the TAWT framework and API of services (abstracting service, keyword extraction service, etc.) a number of application systems and services were implemented [11–15]: a tool for selecting thematic synonyms, a system for searching people by interests in social networks, a plug-in for Google Chrome browser to identify fraudulent messages on the social network VKontakte, a mobile application for extracting and simplifying text data from QR codes TouristHelper 2.0 with support for the Russian language, a set of software tools for analyzing technical documentation.

The set of tools is designed for a wide range of users: researchers in the field of linguistics, developers of automatic text analysis tools, applied applications, in which the use of automatic text analysis tools makes it possible to implement new attractive smart functions.

3. Open Automated Text Processing System for the Russian Language

The system is based on three directions in NLP: linguistic, statistical, analytical. The storage system integrates the processing results of these three directions. It involves the storing of intermediate and final results and text analysis analytic tools. Linguistic and statistical processing tools are sets of operations, individual analysis algorithms and components. The main principles of development and filling the system are accumulation of incoming data by analytical system, constructed in accordance with the adaptive-dynamic model of information transformation [9], this makes the system dynamic and allows expanding the internal "knowledge" of the system. This knowledge is stored as a two-level semantic representation, which includes semantic networks of the upper and lower levels. The upper level is built on the basis of explanatory dictionaries, the lower – on the analyzed text [9].

All basic processing tools are tied to the main stages of automated text processing: graphematic, morphological, syntactic, and semantic. The raw text is the input to the processing tools, the output are the results of statistical and linguistic analysis in the form of various structures: dictionaries, sentence lists, semantic networks, frequency distributions of letters, words, connections between word, etc., all of them are available to a user in the storage system. The resulting structures can represent both the final results of the analysis and also can be used as input for the analytical processing tool – scenario language. This language is an extensible tool focused to maximum ease of use (see figure 2).
It is based on the use of operations executed on the extracted information: sets of words with their morphological characteristics, concepts, sentences, syntactic structures, etc [9]. This allows solving complex problems on a big volume of the source data and process the data on higher levels. The language is designed for users to write their own analysis algorithms; the system also provides typical templates of algorithms for which processing parameters can be set. After executing the scripts, the analysis results are obtained in the form of new structures or numerical values.

Some program tools for external use were created by choosing some of the most popular scenarios that solve practical problems into independent analysis services. Each service of the system has a program interface, a format for passing parameters to run the analysis script in the service, and a format for obtaining the result.

The system is implemented with Java EE tools using the GWT and a set of text analysis libraries: a morphological text analysis library based on the A. A. Zaliznyak's dictionary [16] statistical and linguistic text analysis libraries.

Thus, the system provides a platform for the creation of text processing research tools without the need to write program code. This helps to identify relevant areas in the practical application of algorithms of automated text analysis, implementation and testing of new tools.

4. Text Analysis Framework
The TAWT framework (Tools for Automated Work with Text) is a set of Java libraries for graphematic, morphological and semantic-syntactic analysis [17]. The structure of the framework is shown on figure 3.

The Graphematic Parser tool implements the graphematic stage of text analysis, based on a set of rules implemented by means of regular expressions. The Graphematic Parser tool uses various sets of regular expressions to highlight bear phrases, sentences, paragraphs in order to recognize the necessary boundaries correctly during the analysis.

The JMorfSdk tool (Java Morphological Sdk) implements the morphological stage of text analysis and provides an opportunity to generate a word according to a given initial form and a set of morphological characteristics, based on a modification of the machine-oriented Russian
Figure 3. Structure of the TAWT.

Grammar Dictionary by A. A. Zaliznyak, developed and supported by the OpenCorpora project. A distinctive feature of the tool is word generation mode. You can get a word according to the given morphological characteristics [18].

The MS tool (Morphological Structure) is a tool for loading, storing, processing, and getting a word form and its characteristics, contains a description of all data structures, converts dictionaries used in JMorfSdk from the original format to the format used in the framework.

The GAMA tool (Graphematic and Morphological Analysis) aggregates the methods of graphematic and morphological stages of text analysis, supports the replacement of tools of graphematic and morphological analysis.

The RFC (Rules for Compatibility) tool provides a set of implemented rules for word compatibility. It includes both simple rules, such as the compatibility of a preposition with a noun in specific case; adjective and noun case consistency, number and gender, etc., as well as more complex rules: verbs, verbal adjectives and adverbs are the control word in relation to a noun. If the noun is in the nominative or accusative case, and there is no preposition in front of it this distance between them should not be more than 4 words, etc.

The expanding of this set of rules will be continued. For example, verb management models implementation is planned for the nearest future. This will significantly increase the accuracy of word compatibility processing by taking into account word collocations, not only at the morphological level, but also at the level of a specific word collocations [10].

The AWF tool (Ambiguity Words Filter) contains an RFC rule-based word disambiguation filter, supports replacing the RFC tool with another model that confirms or denies the compatibility of a pair of input words.

The SP tool (Syntactic Parser) implements the semantic-syntactic stage of analysis; it
requires a pre-processed text, where each word has its morphological characteristics. Separate components use the filter to eliminate ambiguity and a set of rules for the words compatibility to establish links between words within the bearing phrase, sentence and paragraph.

The SPN tool (Search Possible Notions) is designed for searching for potential concepts and key phrases based on analysis of dependency networks obtained as a result of the SP tool. The tool is based on the approach that the most common phrases in a large selection of texts written by a person in a natural language can be concepts.

A concept can be from 2 to 17 words length, so all word combinations of length from 2 to 17 words are investigated. Then, all word combinations are counted, including homoforms transposed into the initial form, so that combinations such as myla rami (washed the window frame) and moem ramy (are washing the window frames) are counted as one word combination.

Then frequencies of all combinations are counted, all word combinations that cannot be a concept are removed, finally forming a list of potential concepts. Phrases that contain verbs, or do not contain at least one noun, or ending in and beginning with a preposition (after normalization) are filtered out [10].

The implementation of text analysis algorithms not only takes into account the peculiarities of texts in the Russian language, but is also focused on working with big data: graphematic and morphological analysis algorithms have constant complexity, and semantic-syntactic analysis algorithms have linear complexity (\(O(n)\)). The average speed of morphological analysis of a text is 1 000,000 words per sec [18]. The speed of semantic-syntactic analysis depends on the style of the text and is about 22.2 sentences per sec. for publicist and fiction texts, and 66.7 sentences per sec for scientific texts.

All tools use a common connection scheme that is a standard for the Java platform: a global binary dependency repository is used, the source code, examples and links to the artifacts are shared on GitHub. This allows using the necessary tools only and, if necessary, easily replace them easily or connect new tools to the created software.

5. Web Portal Automated Text Analysis

The Automated Text Analysis portal was created to aggregate and share information on the theoretical foundations of computer linguistics and research results in this area. It provides the opportunity to use research results for a wide range of users. The portal has a number of automatic text analysis services.

Also, the portal contains information on the basics of computer linguistics, a list of popular word processing programs, reference information on the open text processing system, documentation on the use of the TAWT framework and functions of the external program interface.

A separate section of the portal contains automatic text analysis services, which are based on the use of the methods of the internal program interface of the open text processing system and the TAWT framework.

The common statistics section shows the results of accumulated statistics on various frequency characteristics of texts in the Russian language (average and maximum word length, length of sentences, list of new words which are missing in the dictionary, frequency distribution of letters, etc.), which is automatically updated in the system when using services.

The statistical analysis service provides text processing tools on different levels of analysis:

- graphematic:
  - statistics of occurrence of individual letters in general and in a certain position of the word in a set of texts, the maximum length of word;
  - the maximum length of the new word (a word that is not in the A. A. Zaliznyaks morphology dictionary);
• maximum word;
• maximum new word;
• average word length;
• average sentence length;
• number of sentences;
• frequency distribution of word lengths, etc;

• morphological:
  • the number of words in the text;
  • vocabulary size;
  • word frequency distribution;
  • frequency distribution of non-service parts of speech;

• syntactic:
  • frequency distribution of word compatibility;
  • frequency distribution of structures.

The automatic text classification service allows getting classification results and provides the ability to configure the service to work with any subjects. The automatic classification service has two operating options: classification and training. The second version of the service provides the support of creating a hierarchical classifier and linking areas to area numbers in the universal decimal classifier (UDC). The service was tested by developing and implementing a classifier in the field of project management and analysis of its results of the subsequent classification of texts.

The portal provides several services for extracting keywords: with an algorithm implemented in the systems scenario language based on the frequency distribution of nouns in texts; a method for extracting keywords based on markem analysis by A. A. Kretov [19], as well as an integrated method, which contains a number of known and proposed methods and an algorithm for combining them, which gives an improvement in the quality of extracting keywords by 8% to known methods [15].

The automatic summarization service for the Russian language [11] provides an opportunity to choose a method for creating a text summary, indicate the desired summary size as a percentage of the source text volume, and extract keywords out of the text. Implementation of the integrated summarization method allowed improving the quality by an average of 15%-20% compared to existing services.

All these available services have a programming interface and relevant documentation in the web-portal Automated Text Analysis. Thus, the portal provides an opportunity to study the main directions and achievements in the field of computer linguistics, use text analysis services to quickly obtain a result or verify the operation of algorithms in order to subsequently use the provided development tools.

6. The Use of the Toolset for Applied Systems
A number of applied systems were created using the tools of the complex, which contains various automated text processing algorithms.

The service for selecting thematic synonyms provides the ability to get lists of synonyms for the word in the Russian language for each topic area selected by the user.

The FriendsFinder application [12] provides the ability to search for people by interests among the users friends on social networks VKontakte and Facebook and build a chain of friends to the desired person.

The plug-in FraudHunter for the Google Chrome web-browser is designed for monitoring and recognizing fraud messages on the social network VKontakte [13]. The plug-in allows
receiving notifications when suspicious messages arrive before they are read by the user. Now the user can decide to open links or perform actions based on the warning received. The analysis algorithm uses machine learning methods, and vectorization methods are based on computer linguistics.

**TouristHelper 2.0** application: An Android application that allows extracting the text or a web page referenced from a QR code and simplify it (the text size reduce and highlighting the keywords) for quick review of the usefulness of the text for the user [14].

The use of computer linguistics is extremely useful in the workflow systems and, in general, for the preparation of various kinds of documentation: design, engineering, etc. In most cases, formal requirements for documents are given in terms of structure, design, and content. Preparation and updating of the documentation, preparation of documents packages during the software development is a very time-consuming task.

The use of TAWT framework on different levels of text processing can significantly simplify many tasks of preparing technical documentation. Some software tools were developed to automate and simplify the solution of some technical documentation preparation tasks:

- validation the structure of the document for compliance with GOST;
- validation of the section ”Terms and definitions” for all abbreviations used in the document;
- search for similar technical solutions based on technical specifications or a description of operating conditions;
- building a brief document content to speed up the search for suitable documents by a specialist.

The developed cross-platform framework can be useful for developers of text analysis tools for scientific research and developers of Java applications to implement new functions or improve the quality of text processing by applying linguistic analysis methods, as well as developers of automated tools to reduce routine actions when working with different types of documentation.

The methods of the external program interface of text analysis services allow using them in any web applications. In the future, it is planned to implement a REST interface for all the main methods of the TAWT framework.

7. Conclusion

The proposed toolset includes the open automated text processing system for the Russian language, the web portal Automated Text Analysis, a number of NLP services, an open program interface for accessing text analysis tools, and the TAWT framework. The proposed toolset provides the ability to create applications for solving practical problems, implementing text analysis means for mobile applications, taking into account the interests of a wide range of users, and also contributes to the research in the field of computer linguistics and the use of its results in applied software.

The use of linguistic analysis can improve the quality of processing texts in many areas and opens up new possibilities for automating the processing of text documents using accumulated experience in the field of computer linguistics, and the use of linguistic analysis contributes to the appearance of new practical and research tasks in NLP.

References

[1] Jurafsky D and Martin J H 2009 *Speech and Language Processing: An Introduction to Natural Language Processing* (New Jersey: Pearson Prentice-Hall)

[2] Malkovsky M G and Starostin A S 2007 Algoritm sintaksicheskogo analiza, ispolzuyemy v sisteme treeton.(the syntactic algorithm used in treeton system) (Moscow, Russia: RGGU, Dialog) pp 516–524
[3] Bolshakova E I, Klyshinsky E S and Lande D V 2011 *Avtomaticheskaya obrabotka tekstov na estestvennom yazyke (Automated natural language text processing and computer linguistics)* (Moscow, Russia: MIEM)

[4] Official site aot http://aot.ru[Accessed:2019-09-30]

[5] Popov E V 2004 *Obshcheniy s EVM na estestvennom yazyke (Talking to computer in natural language)* (Moscow, Russia: Editorial)

[6] Gate - general architecture for text engineering, the university of sheffield https://gate.ac.uk[Accessed:2019-09-30]

[7] Lingpipe http://alias-i.com[Accessed:2019-09-30]

[8] Uima, apache uima project http://uima.apache.org[Accessed:2019-09-30]

[9] Balakirev N E and Politsyna E V 2013 *Vestnik VGU* 162–168

[10] Porechny A S 2018 Realizacija krossplatformennogo frejmworka dlja avtomaticheskogo analiza estestvenno-jazykovogo teksta (implementation of the cross-platform framework for automatic processing of natural language text) *Proceedings of the XLV Gagarin Science Conference* (Moscow, Russia: MAI) pp 112–3

[11] Official site of automatic text summarization service http://abstracts.textanalysis.ru[Accessed:2019-09-30]

[12] Official site of friend finder application http://friendsfinder.textanalysis.ru[Accessed:2019-09-30]

[13] Petrash Y and Tikhonova D 2019 Razrabotka instrumenta dlja vyjavlenija moshennicheskikh soobshhenij v social’noj seti "vkontakte" (development of the tool for fraud messages identification in the social network "vkontakte") *Computer Science: problems, methodology, technologies* vol 1 (Voronezh, Russia: VSU) pp 1513–18

[14] Official site of friend finder tourist helper 2.0 https://play.google.com/store/apps/details?id=ru.textanalysis.touristhelper[Accessed:2019-09-30]

[15] Ivashchenko M 2018 Analiz metodov avtomatizirovannogo vydeleniya kljuchevykh slov iz tekstov na estestvennom jazyke (analysis of automated keyword extraction methods in natural language texts) *Computer Science: problems, methodology, technologies* vol 6 (Voronezh, Russia: VSU) pp 19–24

[16] Zaliznyak A 2003 *Grammaticheskij slovar’ russkogo yazyka (Grammatical dictionary of the Russian language)* (Moscow, Russia: Russian Dictionaries)

[17] Politsyna E V, Politsyn S A and Porechny A S 2018 The framework for hypothesis verification and analysis of natural language processing for the russian language *Supplementary Proceedings of the Seventh International Conference on Analysis of Images, Social Networks and Texts (AIST-SUP 2018)* (Aachen, Germany: CEUR WorkShop Proceedings) pp 25–33

[18] Politsyna E V, Politsyn S A and Porechny A S 2018 Development of the cross-platform library of morphological analysis of the russian language text for industrial software Proceedings of the Central and Eastern European Software Engineering Conference Russia Moscow (CEE-SECR’18) (New York, NY: ACM)

[19] Kretov A A, Voronina I E, Popova I V and Dudkina L V 2019 *Vestniuk VGU* 1 68–72