Generalized Semantic Analysis Algorithm of Natural Language Texts for Various Functional Style Types

Natalia Sharonova [1], Iryna Kyrychenko [2], Iryna Gruzdo [2] and Glib Tereshchenko [2]

[1] National Technical University "KhPI", Kyrpychova str. 2, Kharkiv, 61002, Ukraine
[2] Kharkiv National University of Radioelectronics, Nauky Ave. 14, Kharkiv, 61166, Ukraine

Abstract

The paper reviews and analyzes existing solutions for determining the meaning of text documents. The issues and problems of determining the meaning of text documents were discussed. The most used models and algorithms of semantic analysis are considered. In the course of the analysis, it was found that when applying classical models of semantic analysis in practice, there is a partial loss of the meaningful meaning of the text, which in turn is not always justified, although it allows you to perform some procedures, but misses a number of features. A theoretical algorithm for the semantic analysis of natural language texts in a generalized form from existing ones was developed. The process of determining the quality of decisions made to find the degree of semantic proximity in text work has been designed.

Keywords

Text fragment, text information, algorithm, text analysis, text meaning determination, method, semantic analysis, latent semantic analysis, library, functional-style type

1. Introduction

At the present stage of development of information technologies, both throughout the world and in Ukraine, tasks related to machine search for information [1] are of relevance and special attention is paid to determining the meaning of text documents and finding dependencies and borrowings.

To date, to determine the meaning of text documents in natural language, there are two large areas that try to solve problems related to the semantic perception of the text, namely the construction of onto-controlled information structures [2] and the definition of the meaning of the text using semantic analysis [3]. Among them, the most successful and effective is the direction associated with the semantic analysis of the text. Considering this observation, the article will be a study of the direction of determining the meaning of text documents using semantic analysis was carried out.

Under the semantic analysis of the text, a component assessment of the number of words or phrases that determine the main meaning of the text (semantic core) and statistical indicators inherent in texts of various types and types will be understood. Considering the functions and style of language and speech, there are six main functional and stylistic types of texts [4]: colloquial, official-business, socially informative, scientific, artistic texts and religious works.

Attention should also be paid to the concept of "language style", since this is what allows you to highlight additional characteristics of the text, thanks to which you can judge not only the authorship, but also the specifics of the construction of various types of texts. According to V. V. Vinogradov, "the style of language is a combination of two factors – "what is said" and "as it is said", that is, it is a purposeful set of linguistic means [4]. At the heart of the concept of language style is the assessment of the relationship of the means of expression to the content expressed".

COLINS-2022: 6th International Conference on Computational Linguistics and Intelligent Systems, May 12–13, 2022, Gliwice, Poland
EMAIL: nvsharonova@ukr.net (N. Sharonova); iryna.kyrychenko@nure.ua (I. Kyrychenko); iryna.gruzdo@nure.ua (I. Gruzdo); hlib.tereshchenko@nure.ua (G. Tereshchenko)
ORCID: 0000-0002-8161-552X (N. Sharonova); 0000-0002-7686-6439 (I. Kyrychenko); 0000-0002-4399-2367 (I. Gruzdo); 0000-0001-8731-2135 (G. Tereshchenko)

When solving the problem of determining the meaning of text documents, it should be remembered that each text is characterized by certain functional and style characteristics, which determine a certain choice of means of analysis at a particular time, and interpretations of the understanding of meaning, which ensure its equivalence. Therefore, one of the subtasks of semantic analysis is the allocation of functional and style characteristics at the stage of pre-processing of the text. To date, there is no single classification of functional styles, many scientists have worked and continue to work on this problem, since this is a very complex and ambiguous issue, both from the point of view of ethics and legal norms.

The semantics of texts of different functional and style types and in different parts of the document are different, but at the same time, depending on the type of text itself, it helps to understand the essence of the text and determine the meaning depending on its specific interpretation within a particular direction. Therefore, in order to more accurately determine the meaning of text documents, in the classics, and search engines perform the process of minimizing the amount of "water" (information that is not the main one but is only used to give the width or volume of the text a certain appearance), but at the same time there is a loss of some characteristics affecting the test. It should also be noted that, performing the procedure of semantic analysis, stop words, in other words, repetitions, also try to remove from the text, since, according to many authors, they complicate the process of determining the semantics of the text. All this affects the process of semantic analysis and, consequently, the result.

The classical semantic model consists of the word itself, its definition, examples of combining it with others and composing phrases, sentences. But in its classical form, it is very difficult to use, because it requires significant computing resources on large volumes of texts [1]. In turn, this imposes a number of restrictions on its use in various fields.

Despite the large number of papers on this subject, the most common are works describing the "bag of words" (or n-gram bag) method [5]. Another commonly used representation method is the Latent Dirichlet distribution (LDA) [3]. Most search giants in their developments use an approach called latent semantic analysis (LSA), or latent-semantic indexing (LSI).).

A relatively new paradigm in the field of semantic analysis is the use of distributed representations for words [6] and for documents [7]. Even though intermediate representations are less readable than earlier methods of determining the semantics of texts, according to some authors, in practice they work quite well and allow you to consider a number of criteria and take into account the presence of "water" and a large number of stop words in the text.

In [5], Mikolov and Chen demonstrate that the paragraph vector method they developed allows semantic analysis to be performed on "large" documents of different meanings. Using representations in the form of vectors that can be used to classify movie reviews or to extract information from web pages. But even the authors themselves note that this technique does not cover different types of work and in practice analyzes only small ones. volume of texts, only annotations, reviews, messages, etc.)

Despite the success of individual studies, very little is known about how well a particular method works in comparison with each other and how sensitive they are to changes in various parameters and characteristics of texts. It is also difficult to understand how the search for texts similar in meaning occurs, what algorithms are available for searching for texts of the same or similar subjects and whether they take into account the functional and style types of text arrays of information. These problems are since the authors do not disclose the algorithm of work and consider their works to be the subject of know-how and due to this, we can say that for 80 years there has been no common solution to the problem associated with determining the meaning of text documents using semantic analysis.

The above circumstances determine the relevance of the task of studying existing algorithmic approaches in the semantic analysis of texts. In turn, the solution of this problem will allow not only to present the process of semantic analysis in the form of steps, but also to develop an approach that will subsequently solve this problem based not only on the language itself, but also on a specific subtask for determining the meaning of text documents.

The article is aimed at analyzing the current state of the problem of determining the meaning of text documents using semantic analysis, reviewing the most used methods and algorithms based on them, as well as formulating and meaningfully formulating the problem of determining the meaning of documents using semantic analysis for different functional and style types.

Statement of the task of the research conducted in the article. There are many algorithms suitable for determining the meaning of text documents using semantic analysis for different functional and style types of texts. In addition, the nomenclature of functional and style types of texts is known.

It is necessary, having formulated several common characteristic criteria on a set of algorithms suitable for determining the meaning of text documents, and based on known algorithms to formulate the task of semantic analysis, taking into account the subtask of determining functional and style types of texts.

The solution of this task involves the implementation of the following sequence of steps:

- clarification of the steps of the text processing process during semantic analysis and, based on the refined algorithm, the formulation of the generalized algorithm.
- meaningful formulation of the task of determining the meaning of text documents using semantic analysis for different functional and style types of texts.
- definition of criteria for assessing the quality of decisions made to determine the fact of semantic proximity of text documents, considering paraphrasing.

As a result of solving the problem, the current state of the problem of semantic analysis of natural language texts will be reviewed, the most used algorithms for semantic analysis will be considered, and a generalized algorithm for semantic analysis of natural language texts will be developed. The process of determining the quality of decisions made to find the degree of semantic proximity in text work has been designed.

2. The state of the problem of semantic analysis of natural language texts

The application of existing methods and algorithms of semantic analysis depends on the application area of the problem being solved, to which functional and style type it belongs and the existing directions in the processing of text information.

There is a wide range of search engines for determining the similarity of the text, namely [8]: utilities for statistical analysis of the text; linguistic technologies and systems; utilities of linguistic analysis of the text (morphology, syntax); natural language processing systems.

During the study in the work [1], the class of natural language processing systems was considered, and the methods and algorithms that underlie them were analyzed. In this class of problems, there are two groups of methods that allow solving the problem of determining the meaning of text documents [1]:

- methods of semantic processing of texts, which are aimed at the so-called "linguistic transformations", namely translation into a foreign language in two directions, both into the language of interest and back to the original; a brief retelling of texts; notes; abstract submission; annotation.
- methods of semantic text processing based on artificial intelligence. This group focuses on "extracting knowledge" from the texts being analyzed, namely the classification of text messages, answering questions, contextual translation and understanding of textual information based on what has been presented. In this class of problems, in most cases, conceptual analysis of texts is used. Within the framework of this, special attention is paid to such problems as the synthesis of knowledge representation systems and the development of systems for semantic analysis and machine "understanding" of texts using ontologies.

In most cases, the following models of semantic text processing are used to determine the meaning of text documents using semantic analysis: "linguistic transformations" over texts [9–11], content analysis [12–18], context analysis [13, 15–17, 19-20], paragraph vector model [21], as well as associative semantic analysis [22].

There are many publications that allow you to perform the so-called "linguistic transformations" on texts, but the most significant contribution to its solution was made by A. Palagin [8]. In accordance with the decision of A. Palagin [8], the typical process of text processing in semantic analysis consists of the following steps:

1. Removal of "stop words" are words that occur in every text and do not carry a semantic load, it is, first, all conjunctions, particles, prepositions, and many other words).
2. Stemming is the process of finding the basis of a word for a given source word. The basis of the word does not necessarily coincide with the morphological root of the word.
3. Application of the algorithm that directly makes up the semantic core.

Another important part is the basic semantic template [10, 11], which will allow you to analyze textual information. BSMS are called the rule according to which there is a semantic dependence in the analyzed text. In accordance with the works of A. V. Mochalov BSMS consists of four main parts [10]:

1. sequences of words or indivisible semantic units for which their morphological features are indicated (in some cases, when it is especially important for semantic analysis, the names of these words and semantic units are given).
2. the names of the semantic relation, which should be formed if the sequence described in the previous paragraph is found in the text.
3. a sequence of numbers that determine the positions in the sequence from clause 1, the elements of which should be added to the queue with priority, according to which words will subsequently be removed from the analyzed sentence fed to the semantic analyzer.
4. the number denoting the priority value, the group of semantic dependencies to which the semantic relation belongs.

According to this algorithm, sentences are received at the input, and at the output the program provides a set of semantic relations formed from the analyzed text.

After the formation of a set of indivisible semantic units, a morphological analysis of each such unit is carried out. Then there is a search for such complex language constructions as introductory, participle and participle turn, subordinate clauses, etc.

Further, the coincidences of each BSMS from one set to another set are sequentially searched, while both for the BSMS and for all sentences their morphological characteristics are considered.

If a match is detected, with some subsets of the set of sentences, for all different subsets and sentences coinciding with the BSMS, semantic dependencies are formed, which are recorded in the database if they are detected in the analyzed text for the first time. The search for BSMS in the offer takes place until all the templates are checked for a match.

Next, the search for the BSMS among the remaining indivisible semantic units of the set in the text is repeated. This continues until all the semantic dependencies described by the BSSS are found in the sentence under analysis.

Taking into account the above, it can be made a platoon that the algorithm proposed by A. V. Mochalov for finding semantic dependencies [10, 11] with the help of BSMS is very complex, and due to its adaptation to the task of detecting dependencies in texts of different thematic orientation in this form, is resource intensive.

Within the framework of the second group of determining the meaning of text documents, the direction of content analysis should be noted. To date, there is no unambiguous definition of what content analysis is, and there are many interpretations of it.

Within the framework of this study, the following definition of content analysis was chosen - a method for identifying and evaluating the specific characteristics of texts and other information carriers, in which, in accordance with the objectives of the study, certain semantic units of content and forms of information are allocated [15].

Content analysis makes it possible to identify individual characteristics of information, as well as their interrelations. In the process of extracting information, the frequency and volume of references to these units in a certain set of texts or other information are systematically measured.

There are several classifications of types of content analysis. The most famous was proposed by R. Merton [15]:

1. Character Counting – A simple count of certain keywords.
2. Classification of symbols in relation – a balance of positive and negative statements about the object of research. It is used to analyze the effective arrangement of symbols for propaganda, to detect contrasting, contradictory judgments and to determine the intentions of the communicator.
3. Analysis by elements – the choice of the main and secondary parts of the text, the definition of topics related to the main and peripheral interests of the audience.
4. Thematic analysis – identification of obvious and hidden topics.
5. Structural analysis – clarification of the nature of the ratio of different materials: complementary, combined, colliding.
6. Analysis of the relationship of various materials – a combination of structural analysis with the study of the sequence of publication of materials, the volume and time of their publication.

Such many types of content analysis is due to the ambiguity of the criteria, characteristics or elements of the content, in relation to which the counting procedure is applied, since these can be individual words, phrases, sentences, paragraphs, texts [23].

With the help of a frequency analysis of the lexical composition of the text, it is possible to determine the theme of the work, to identify specific lexemes inherent in the text, to establish the style of the text, to highlight the genres of the text, etc. Depending on the frequency of significant words in the text, one can judge the semantic dominants of the text, and by the most frequent words it is possible to determine the semantic accents of the text.

The generalized algorithm of content analysis based on the scheme from [19] is as follows:

Step 1. Preparation of the research program

1. Formation of the goal, objectives, hypothesis of the study.
2. Selection of the time interval, selection of the source of analysis, selection of categories and indicators of evaluation and compilation of the dictionary.
3. Systematization of information on the problematic feature of determining the sample size.

Step 2. Collection of information

1. Correlation of categories and subcategories of content analysis with specific content in the text.
2. Encoding information, recording the frequency and volume of mentions of categories and subcategories of content analysis.

Step 3. Analysis of the results obtained

1. Statistical processing of the obtained quantitative data.
2. Interpretation and visualization of the obtained data based on the obtained tasks.

It should be noted that a simple frequency count is not an exhaustive or sufficient criterion in determining the meaning of text documents. Also, another problem is the comparison of texts of different lengths, since it is necessary to compare not simple frequencies, but conditional ones, i.e., shares that constitute a certain category in the first and second text.

Algorithm of content analysis according to [24]:

1. Compilation of a table of content analysis (category + unit of analysis).
2. Calculation of the frequency of use of keywords in messages (encoding matrix).
3. Calculation of the share of keyword usage in messages (encoding matrix).
4. Context consideration.

- Opinion of arbitrators (coders).
- Consideration of the topic (a certain combination of words or concepts, embodied, for example, in a phrase).

5. Is the mention given in a positive or negative sense?

- The Q-sort method.

6. Janis's formula.

According to this algorithm, the same category can be expressed by different units of analysis: word, topic, paragraph, judgment, hero, social, message as a whole (book, letter, speech, newspaper). That is, it can be concluded that a significant problem is the choice of the assessment itself, in what exactly it is necessary to evaluate in words, sentences, in percentage content, etc.

An integral part of content analysis is contextual analysis. The main essence of contextual analysis is that not the entire text is analyzed, but only a certain sample of the object of analysis from it, which is the context of the use of a certain characteristic [20]. It should be noted that there are many ways to set the context. Contextual analysis will allow you to evaluate not only for individual categories of the analyzed text, but also for interrelations.

It is possible to introduce logical relations of jointness, contradiction, subordination, etc. on a set of vectors [25]. Thus, a certain logical model of the subject area referred to in the text is set, or a model of the cognitive map inherent in the author of the text.

Context analysis algorithm [26]:

1. Determine the totality of sources or messages to be studied using a set of specified criteria that each message must meet.
2. Formation of a sample set of messages.
3. Identification of units of analysis (words or topics).

4. Allocation of units of account, which may coincide with semantic units or be of a specific nature.
5. The counting procedure itself. It is generally like the standard methods of classification by selected groupings.
6. Interpretation of the obtained results in accordance with the goals and objectives of a particular study.

The main problem is that it is necessary to anticipate not only the mentions that may occur, but also the elements of their contextual use, while a detailed system of rules for assessing each case of use should be developed.

Another model of semantic text processing is the paragraph vector model, which is described in the work [20], the authors called it "Distributed Memory". This model, when determining the meaning of text documents, adds a memory vector to a standard language model aimed at determining the topic of the document. In general, the algorithm of paragraph vectors for semantic text processing is as follows:

1. a dictionary is created
2. the corpus is read, and the occurrence of each word in the corpus is calculated
3. an array of words is sorted by frequency and rare words are removed
4. a Huffman tree is built – to encode the dictionary to reduce the computational and temporal complexity of the algorithm
5. the so-called sub-proposal is read from the corpus, which is the basic element of the corpus – a sentence, a paragraph, an article, after which the most frequent words are removed from the analysis (sampling)
6. going through the sub-proposal and calculating the maximum distance between the current and predicted word in the sentence
7. a direct propagation neural network with the activation function of hierarchical softmax and / or negative sampling is used
8. the vector representation of the words is calculated. The vector representation is based on contextual proximity, the essence of which is that the words encountered in the text will have close coordinates of the vector-words.

A paragraph vector is concatenated or averaged using local contextual word vectors to predict the next word.

It should be noted that in the process of analyzing a paragraph vector, you can further simplify if you do not use the local context in the prediction task. With this approach, the parameters of the classifier and word vectors are not used, and the backpropagation algorithm is used to adjust the parameters of paragraph vectors.

In practice, the paragraph vector model is effective for the task of finding semantic similarity for large chunks of text. The article [21] uses an example to prove that the paragraph vector model is superior to the LDA in finding semantic similarity in Wikipedia articles at different vector sizes, but it is surprising that vector operations can be performed on articles like word vectors.

Associatively, semantic analysis [22] is a set of lexemes denoting concepts united by cause-and-effect relationships and at the same time having identical schemes in the semantic structure that are in a special kind of relationship. In accordance with the conclusion of the authors [22] can be divided into three stages:

- Transition from words and phrases of sentences to the corresponding semantic meanings – concepts of ontology.

1. Search for the meaning of a word from a variety of possible alternatives to concepts, which is semantically the closest to the meanings of neighbor words from the local environment of this word.
2. Finding the degree of proximity of the shortest path between concepts in the network of ontology. Calculation of the distance in ontology between the assumed meaning of the word and the conceptual meanings of the words of the immediate environment.
3. Defines in the semantic network of the ontological knowledge base the concept corresponding to the correct meaning of the word or phrase in the text.

- Assembling semantic frames of text sentences.

1. The choice of the type of slot to fill with the meaning of the concept of a word depends on the syntactic position of this word in the grammatical structure of the sentence.

2. Filling in the slots of the frame structure of the sentence, analyzing the sentence parse tree and the syntactic positions of words and phrases for each concept using semantic-syntactic tables of modal-role cases like File more cases.
3. Construct a semantic frame of the current input text sentence.

- Unification of semantic structures of text sentences into a single semantic network of text. Combining two structures into one network is performed on the principle of combining semantically identical vertices, that is, if there are vertices in the structures *G1* and *G2* that refer to one semantic concept, these vertices are combined into one. At the output of the system, a semantic network of input text is generated, which contains in the vertices the concepts of the text connected by arcs of semantic relations.

The use of the considered models in practice in most cases leads to a partial loss of the meaningful meaning of the text, although it makes it possible to perform semantic processing of texts and allows you to group documents by formal features more quickly. Therefore, it is necessary to pay special attention to the methods of extracting data from natural language texts, as they allow you to perform preliminary preparation of texts during semantic analysis and allow you to understand why there is a decrease in accuracy in the process of semantic analysis of the document.

3. Setting the task of determining the meaning of text documents using semantic analysis for different functional and style types of texts

Based on the considered concepts and existing solutions, a meaningful formulation of the task of determining the meaning of text documents using semantic analysis for different functional and style types of texts can be formulated.

The initial data of the task is a lot of texts of different functional and style types and of different volumes.

While solving the problem, it is necessary to obtain data on the degree of semantic proximity of textual work, considering possible facts of borrowing or paraphrasing.

Because the results of the problem will be used to determine the meaning of text documents when calculating the degree of semantic proximity of text work, taking into account possible facts of borrowing or paraphrasing, it is advisable to introduce the following requirements:

- search should be implemented for text different functional and style types.
- search should be implemented for different volumes of texts.
- it should be possible to choose a search tool.
- the search should consider possible errors and typos in the analyzed text.
- for each word from the text, the setting of many synonyms should be provided.
- a set of all documents must be found – the degree of proximity of which is close.
- an analysis of the results should be provided.
- there should be a mode for re-calculating the degree of semantic proximity with a change in the order of the components of the attribute group and a change in the functional and style types of texts.

The specificity of this task involves the development of a special method that will make it possible to consider all the above requirements as fully as possible. The solution of the problem is complicated by the fact that today there is no single classification of functional and style types of texts, but such a classification will allow us to identify additional characteristics of the text, thanks to which it is possible to judge not only the authorship, but also the specifics of the construction of various types of texts that also differ in volume. It is also necessary to feed the repetitions, and not removing them from the text, since they affect the process of semantic analysis and, therefore, the result.

These features make it possible to develop a specialized methodology for determining the meaning of text documents using semantic analysis for different functional and style types of texts. Which will allow you to find the semantic proximity of different more accurately at first glance types of documents.

A generalized algorithm for semantic analysis of natural language texts for various functional and style types can be represented as:

1. selection of the source of analysis, selection of evaluation indicators
2. systematization of information by the selected feature relative to the volume of the database

3. definition of the functional and style type of text for the analyzed text (style, substyle, genre style, genre substyle). Define the characteristics of functional and style types for the text being analyzed
4. checking whether there are significant differences between groups for the analyzed text in terms of the selected functions or variables
5. assignment of the document to one of the specified groups of functional and style types
6. determination of the direction of semantic analysis
7. definition of stylistic techniques and other, stylistically marked means of language and text segments
8. the timing of the algorithm that directly constitutes the semantic core (e.g., latent semantic analysis or paragraph vectors, etc., etc.)
8.1. search for the meaning of a word from a variety of possible alternatives to concepts, which is semantically the closest to the meanings of neighbor words from the local environment of this word.
8.2. definition of dependencies between words
8.3. finding the degree of proximity between elements (by sentences, phrases in the text, words)
8.4. assembly of semantic frames of text sentences
8.5. constructing a semantic frame of the current sentence of the input text
8.6. unification of semantic structures of text sentences into a single system of the analyzed text
9. analysis by elements (selection of the main and secondary parts of the text, determination of topics related to the main and peripheral interests of the audience)
10. thematic analysis (identification of obvious and hidden topics)
11. analysis of the relationship of various materials (combination of structural analysis with the study of the sequence of publication of materials, the volume and time of their publication)
11.1. correlation of categories and subcategories of content analysis with specific content in the text
11.2. finding matches of each basic semantic pattern from one set to another set
11.3. encoding of information, recording the frequency and volume of mentions of categories and subcategories
12. issuance of values of sets of semantic relations formed from the analyzed text
13. finding all possible semantic dependencies, in other texts with which the original one is compared

It should be noted that this is a theoretical algorithm that is derived from the totality of existing algorithms considered in the paragraph "State of the problem of semantic analysis of natural language texts" of this article. Therefore, it is necessary to perform its verification, for this it is necessary to perform its programmatic implementation. It should also be noted that most likely within the framework of testing, these algorithms will be refined and modified, so we can say that this is the first version of the algorithm for semantic analysis of natural language texts. It is also necessary to determine how the degree of semantic proximity of texts will be evaluated.

4. Determining the quality of decisions made to find the degree of semantic proximity of texts

Assessment of the degree of semantic proximity of texts. The task of determining the degree of semantic proximity of texts is to experimentally study and assess the adequacy of the semantic analysis model, as well as to decide on the proximity of texts depending on the functional and style types of text.

Let the overall decision-making task S be defined formally [27] by the Quartet:

$$S = (\Theta, U, L, P), \tag{1}$$

where Θ – is the set of characteristics of the decision-maker (LPR), i.e., the set of possible situations on the set of possible results of observations; $X = \{X_1, X_2, \dots, X_n\}$;
$U = \{u\}$ – many possible solutions;
$L: \Theta \times U \to R$ – limited actual loss function; $L(\Theta, U)$;
P – is some statistical regularity on Θ, that is, a non-empty family of finite-additive probable measures on 2^{Θ}, or a given class of functional regularities closed in the corresponding topology.

It is required to choose one $u \in U$ that would minimize losses under the unknown $\theta \in \Theta$. Note that in our case of solving unstructured problems, it is possible to extract knowledge with the help of experts, learn from scenario examples and form functional patterns on $\Theta.$

This allows you to define the decision-making scheme (DSM) Z as an ordered triple:

$$Z = (\Theta, U, L) \in \boldsymbol{Z}. \tag{2}$$

from class Z of the permissible variants of solutions formed by the SPPR allowing to realize the algorithmic degree of semantic proximity of the texts P in the form for deriving the desired solutions.

Therefore, the formula (2) can be clarified: the SPPR $Z = (\Theta, U, L) \in \boldsymbol{Z}$ and the class $P(\Theta)$ of regularities characterizing $\{\theta_n\} \in \Theta$.

It is necessary to choose a sequence of solutions $\{u_n\} \in U$ so that the average losses (risk *R(S)*) are minimal for a given class of patterns $P(\Theta)$.

So, in general, the adequacy and effectiveness of the proposed method for assessing the degree of semantic proximity of texts will be calculated using the minimax risk criterion R. The optimal solution will be considered a solution u^{**} for which the condition is met:

$$\underset{\theta}{sup} R(\theta, u^{**}) = \underset{u}{inf}\, \underset{\theta}{sup} R(\theta, u) \tag{3}$$

In practice, the risk R is estimated by the amount of error that the found decisive rule makes on control situations $\{\theta_k, k \in K\} \subseteq \Theta$. The error value is characterized by the ratio of the number of decisions incorrectly made by the rule to the total number of decisions presented. The above assessment technique will be used to conduct an experiment to determine the degree of semantic proximity of texts. Experimental assessment of the adequacy and effectiveness of the developed algorithm of semantic analysis Natural language texts for various functional and style types will be conducted on specially developed software.

5. Conclusions

In the course of this work:

- an analysis of the current state of the problem of determining the meaning of text documents was carried out;
- the importance of semantic analysis was determined;
- an overview of the most used algorithms of semantic analysis is carried out;
- the steps of the text processing process in semantic analysis have been specified;
- a generalized algorithm for determining the meaning of text documents using semantic analysis for different functional and style types of texts has been developed;
- a meaningful formulation of the task of determining the meaning of text documents for different functional and style types of texts was carried out;
- the criteria for assessing the quality of decisions made to determine the fact of semantic proximity of texts, taking into account paraphrasing, have been determined.

As a conclusion throughout the work, it can be said that semantic analysis has a high practical application, determining the meaning of text documents.

It should be noted that today there is no single classification of functional styles, but the solution of this subtask will allow you to operate with the characteristics of various style texts and thereby increase the accuracy of determining the meaning of text documents. Therefore, it is necessary to pay attention to its solution in the future.

The results obtained will make it possible to continue work on the analysis of texts for the presence of textal borrowings and borrowings of the idea in them, as well as in determining the authorship of the text, taking into account its paraphrasing. At the moment, work is underway to create a system for assessing the degree of semantic proximity of texts. Within the framework of this system, the developed algorithm is used.

6. References

[1] G. Tereshchenko, I. Gruzdo. Overviewand Analysis of Existing Decisions of Determining the Meaning of Text Documents, 2018 International Scientific-Practical Conference Problems of

Infocommunications. Science and Technology PIC S&T'2018, 2018, IEEE, Kharkiv, Ukraine p.645-653 doi: 10.1109/INFOCOMMST.2018.8632014.

[2] O. Ye. Stryzhak, V. V. Gorborukov, O. V. Franchuk and M. A. Popova. “Ontologiya zadachi vyboru ta yiyi zastosuvannya pry analizi limnologichnyx system [Ontology of the problem of choice and application in the analysis of limnological systems].” Ekologichna bezpeka ta pryrodokorystuvannya: Zb. nauk. pracz M-vo osvity i nauky Ukrayiny, Kiyiv. nacz. un-t bud-va i arxit., NAN Ukrayiny, In-t telekomunikacij i global. inform. prostoru; redkol.: O. S. Voloshkina, O.M. Trofymchuk (golov. red.) [ta in.]. Kyev, Ukraine: no.15, pp.172-183, 2014.

[3] D. Blei, M. Jordan and A. Ng, Latent Dirichlet Allocation, Journal of Machine Learning Research, vol. 3, no. 1, pp. 993-1022, 2003.

[4] S.M. Shcherbina, Scientific style in the system of stylistic differentiation of modern literary languages, Scientific Journal of the National Pedagogical University named after M.P. Drahomanov. Series 9: Current trends in language development, 2017, Vol. 16, pp. 261-268. URL: http://enpuir.npu.edu.ua/handle/123456789/20001.

[5] T. Mikolov, K. Chen, G. Corrado and J. Dean, Efficient estimation of word representations in vector space, Arxiv.org, 2013. URL: https://arxiv.org/abs/1301.3781.

[6] Q. V. Le and T. Mikolov, Distributed representations of sentences and documents, International Conference in Machine Learning, Arxiv.org, 2014. URL: https://arxiv.org/abs/1405.4053.

[7] Z. Harris, Distributed structure, Word, vol. 10, no. 2-3, 1954, pp. 146-162.

[8] I. V. Gruzdo. Problemi analiza estestvenno-yazikovix tekstov dlya obnaruzheniya plagiata v uchebnix rabotax. [Problems of the analysis of natural language texts for detecting plagiarism in educational work]. Radioelektronni i komp'yuterni systemy, no.1 (49), 2011, pp.130-138.

[9] K. Palagin, S. Kryvy, V. Velichko and N. Petrenko, K analizu estestvenno-jazykovyh ob’ektov [To the analysis of natural language objects], International Book Series. Supplement to the International Journal Information Technologies & Knowledge, ITHEA, Sofia, Bulgaria vol. 3, no. 9, 2009, pp. 36-43.

[10] A. V. Mochalov, Algoritm semanticheskogo analiza teksta, osnovannyj na bazovyh semanticheskih shablonah s udaleniem [Algorithm for the semantic analysis of text, based on the basic semantic templates with deletion]. Nauchno-tehnicheskij vestnik informacionnyh tehnologij, mehaniki i optiki, no.5, 2014, pp.126-132.

[11] V. A. Kuznetsov, V. A. Mochalov and A. V. Mochalova, Ontological-semantic text analysis and the question answering system using data from ontology, in ICACT Transactions on Advanced Communications Technology (TACT), vol. 4, issue 4, July 2015, pp. 651-658.

[12] V. Shalak. Jelementy matematicheskih metodov komp'juternogo kontent-analiza tekstov [Elements of mathematical methods of computer content-analysis of texts], Vaal.ru, URL: http://www.vaal.ru/show.php?id=146.

[13] Pingali, K., Bilardi, G. (2015). A Graphical Model for Context-Free Grammar Parsing. In: Franke, B. (eds) Compiler Construction. CC 2015. Lecture Notes in Computer Science , vol 9031. Springer, Berlin, Heidelberg. doi: 10.1007/978-3-662-46663-6_1.

[14] N.V. Kostenko, V.F. Ivanov. Experience of content analysis: Models and practices, Monograph. Free Press Center, Kyiv, 2003.

[15] Technologies of document analysis in sociological research. THEME 5. Formalized analysis of documents (content analysis), 2018. URL: https://bookonlime.ru/lecture/tema-5-formalizovannyy-analiz-dokumentov-kontent-analiz.

[16] O.V. Ivanov, Quantitative content analysis: the problem of context, Bulletin of V.N. Karazin Kharkiv National University. Series: Sociological research of modern society: methodology, theory, methods, vol. 30, Kharkiv National University named after V.N. Karazina, Kharkiv, 2012. № 999. 95–99.

[17] Q. Yang, A novel recommendation system based on semantics and context awareness, Computing, 2018, vol. 100, no. 8, pp. 809-823.

[18] Fei Long and Hongju Cheng. Improved Personalized Recommendation Algorithm Based on Context-Aware in Mobile Computing Environment. Wirel. Commun. Mob. Comput. Vol. 2020:1-10. doi: 10.1155/2020/8857576.

[19] R. I. Loktev and S.M. Zuev Kontent-analiz kak metod issledovanija osobennostej zhiznedejatel'nosti postojannogo naselenija pribrezhnyh naselennyh punktov JaNAO [Content

analysis as a method of studying the features of vital activity of a permanent population of coastal settlements of the Yamal-Nenets Autonomous Okrug]. Arktika i Sever, no.26, pp.126-135, 2017.

[20] V. Maran, A. Machado, G. M. Machado, I. Augustin, and J. P. M. de Oliveira, “Domain content querying using ontology-based context-awareness in information systems”, Data & Knowledge Engineering, vol. 115, 2018, pp. 152–173.

[21] A. Dai, C. Olah and Q. Le, "Document Embedding with Paragraph Vectors", Arxiv.org, 2018. URL: https://arxiv.org/abs/1507.07998.

[22] G. Zav'jalov, D. Mal'cev, Ju. Solodovnikova, R. Jaremchuk, Kontent-analiz [Content Analysis]. URL: http://www.myshared.ru/slide/909850/.

[23] I. Shubin, S. Solonska, S. Snisar, V. Slavhorodskyi, V. Skovorodnikova, Efficiency Evaluation for Radar Signal Processing on the Basis of Spectral-Semantic Model. Proceedings of the 15th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering, TCSET 2020, 2020, pp. 171–174.

[24] A. A. Marchenko, A. A. Nikonenko, Kontekstnyj semanticheskij analiz teksta. Sistema tekstovogo monitoringa i kachestvennogo ocenivanija fokusnogo obekta [Contextual semantic text analysis. The system of text monitoring and qualitative assessment of the focal object]. Shtuchnij intelekt. 2008. No 3. pp. 808-813. URL: http://dspace.nbuv.gov.ua/bitstream/handle/123456789/7155/02-Marchenko.pdf?sequence=1.

[25] K. Smelyakov, S. Smelyakov, A. Chupryna. Advances in Spatio-Temporal Segmentation of Visual Data. Chapter 1. Adaptive Edge Detection Models and Algorithms. Series Studies in Computational Intelligence (SCI), Vol. 876. Publisher Springer, Cham, 2020. pp. 1-51. doi: 10.1007/978-3-030-35480-0.

[26] Kontent-analiz [Content Analysis]. Wikipedia. URL: https://ru.wikipedia.org/wiki/%D0%9A%D0%BE%D0%BD%D1%82%D0%B5%D0%BD%D1%82-%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7.

[27] B. M. Konorev, V. S. Harchenko, G. N. Chertkov, Instrumental'naja sistema dlja podderzhki jekspertizy i nezavisimoj verifikacii kriticheskogo PO: principy postroenija i primenenija [Tool system to support the examination and independent verification of critical software: principles of construction and application]. Informacionnye tehnologii i bezopasnost'. Kyiv. NANU, 2003. №4. pp. 85-91.