Texts Segmentation and Semantic Comparison: Method and Results of its Application

Yury Rogozov, José Gregorio Bermúdez Soto, Aleksandr Sviridov, Yulia Lipko, Svetlana Belikova, Aleksandr Belikov, Sergey Kucherov, Oksana Shevchenko, Elena Borisova, Aleksandr Egorov
Southern Federal University
Taganrog, Russian Federation
{yrogozov, sbermudes, asviridov, ylipko, belousova, anbelikov, skucherov, ovshevchenko, eborisova, egorov}@sfedu.ru

***Abstract*—The paper describes the approach of scientific texts comparing. The approach is characterized by the comparison of significant textual passages that consist of elements of meaning. The proposed method of scientific texts comparison differs from others known methods by the using of segmentation with semantic criteria, taking into account synonyms. That allows to automatically detect the semantic similarity between two compared texts, and take into account both the morphological structure of the text and its lexico-semantic content. The results of practical application of the method are also presented in the paper.**

I. Introduction

All intelligent systems of text processing in natural language, at the present stage, have found their application in various fields and solve specific problems. Systems that use automatic text processing are, for example: publishing systems; automated lexicographic systems for dictionaries preparation and use; information retrieval systems; machine translation systems; systems of understanding and speech recognition and others.

The tasks of textual information processing are characterized by different set of input data and the required form of its presentation, target result and proposed approaches. The results obtained in this field and their practical implementation show that some of these problems require further research. One of these tasks is the comparison of texts, namely the task of identifying the semantic proximity of two texts in natural language.

The problem of automatic processing of natural language during the detection of semantic proximity is that different languages have different semantic and grammatical features, while existing algorithms are successfully used only for one single language processing. Overcoming this problem requires the creation of means for building the semantic constructions of natural language, which at the moment is not a solved problem.

For the computer processing of texts in natural language, first of all it is necessary to solve the problem of creating means for converting a language (for example, Russian, Spanish, English, etc.) into a formalized language similar to programming language. General principles for creating tools for word processing systems include the following components: fragmentation or separation, morphological analysis, syntactic and semantic analysis, and the output of one component is the input for the next. It should be noted that the existing approaches and methods do not consider **the meaning of the text** as a criterion for texts comparison [1], which in this paper, unlike known ones, means a text passage that does not contain anaphoric references associated with the words of another text passage containing at least one verb, the type and category of which expresses the action.

On the basis of the definition of text meaning proposed above, the approach of comparing scientific texts was formulated. On the basis of this approach the method and algorithms for semantic comparison of scientific texts are developed. This method and algorithms can be used in applications of automatic detection of plagiarism to improve the detection effectiveness.

II. Related work

A. Texts Segmentation methods

The method of arbitrary text passages extracting [20] as a basic unit of division uses an arbitrary text passage instead of a word. Arbitrary text passage is any continuous sequence of text in a document. There are many ways to divide or structure a document into text passages. As a rule, passages can be formed using a fixed number of phrases or sets of phrases split up from each other by separators, such as a dot or a stop word (pronouns, prepositions, etc.). In particular, in [20], it is proposed to extract arbitrary passages based on the function (word, number of words, etc.) of a search engine query and an indexing system.

The TextTiling algorithm [21] is a segmentation algorithm that aims at identifying substructures in documents. The algorithm divides explanatory texts into units of speech from a set of points. Unlike many speech models deal with hierarchical segmentation, here the text is presented as a linear sequence of segments. The algorithm consists of three main parts: preprocessing, calculating a lexical score and identifying limits. In the first part, stop words or negative words (prepositions, articles, etc.) are omitted, a morphological analysis of the text is performed and documents are divided into sequences of significant words, regardless of punctuation marks; these sequences are called sentences. Then there denotes the lexical combing points for the space between groups of sentences by two methods. The first method compares neighboring text blocks formed by a group of sentences and assigns an estimate of the similarity between these blocks by the number of common words. The second method, called the vocabulary insertion, forms text intervals with sentences and assigns a lexical score at the midpoint of the range, based on the number of new words (words that have not been found in the text before) around this midpoint. And

finally, the identification limit is performed identically for the two methods of lexical score.

The algorithm "Fragmentation by Dynamic Programming" is a hashing algorithm using dynamic programming [23], belongs to the sub-text segmentation approach, but, unlike TextTiling, offers a method that applies a window that runs throughout the text and defines each paragraph for the most similar item in the window. This forms a series of upper and lower points for analysis. This method is very useful when you need to control the length (by the number of words) of the segments.

This segmentation method uses a dynamic programming method to detect the boundary of the minimum cost segment compared to the lexical cohesion curve [22] between points, the preferred length of user-defined segments, and the parametric function defined by the cost length. The logic followed by the algorithm is the determination of the segmentation value for each item sequentially from the first to the last. In addition, each item is determined by its limit, which will be the last paragraph of the previous segment.

The method of fragmentation by dynamic programming is able to determine the optimal correspondence between the length of the received segments, the required length for them and the similarity value associated with each item. There is a disadvantage that the document cohesion vector associates each item with the highest value of similarity in their window, but does not take into account that this value can correspond to the above or lower paragraph. Despite this, the specified value determines whether the associated paragraph includes the last processed segment or not, and this algorithm instantly extends to the next item. As you can see, a situation where high similarity with the preceding paragraphs leads to inclusion in the lower segment, that weakens the value of the method.

The last two methods [21], [23] eliminate stop words or negative words (prepositions, articles, etc.). This represents a disadvantage for solving the problem, since these words provide the meaning of the segments.

B. Semantic comparison methods

Currently, the search for similarities between texts has a great practical application, including the detection of plagiarism in scientific and creative research. In [5], three main categories for detecting textual similarity are mentioned: word-based comparison, linear point-based search used by search engines and stylistic analysis. There are also methods based on various characteristics of texts, for example, methods based on semantics for detecting plagiarism [6], [7] and for information search, as shown in [8].

Many of the developed and applied models for determining the degree of texts semantic similarity, mainly used the emphases in search of characteristics that coincide in both texts, thus ensuring the discovery of whether the two texts have a similar meaning [9].

In the study [8], an attempt was made to create a semantic model for revealing the implicit arguments in texts. The authors say that, even though this is an easy task for the human reader, it is difficult for computers, because there is no way to tell them that the argument can be output several times in the text.

The essence of the textual semantic similarity is that the meaning of the two texts should be similar. This concept is broader than to find the degree of textual similarity, as in the case of the above algorithms, which measure only the number of terms of both texts, but do not measure the similarity of these texts. Moreover, the similarity must be expressed by a specific value.

In general, the concept of information relevance is based on its quantitative assessment. Schematically, this can be represented as follows (1): there is a document D and a query Q, the ultimate goal is to measure the similarity or relevance between them.

$$sim(D, Q)=? \quad (1)$$

In order to determine this relevance, information retrieval systems directly apply a number of functions that are called similarity measures that quantify the relevance between a document and a query.

These measures are based on the number of terms that are jointly found in both the document and the query.

Thus, for the proposed method, unlike search engines, the ultimate goal will be to determine the similarity or relevance between the two texts *sim (D1, D2).*

So, it is possible to calculate the similarity between documents D1 and D2 in the similarity function between the corresponding text passages as in the equation (2):

$$sim(D1,D2)= f(Xij .. Xnm) \quad (2)$$

As noted above, there are also methods that calculate the similarity between two documents through algorithms based on the frequency of terms occurrence within both documents, but these are nothing more than methods that apply to comparing a query and a document in which one of them is a query. Thus, it is enough to check the similarity calculation between the request and the document.

A measure of similarity makes it possible to determine the similarity between two text segments (whether it's a document or a passage) and a query, or in our case between two text passages. Traditionally, these measures are based mainly on terms that exist in both texts and in the request, and also on the discriminatory meaning of each term.

Information search methods implement these similarity measurements, defining document D as a set of pairs of values (di, ni), in which di is the term, ni is the number of repetitions of the specified term in the document. The value of N represents the size of the document in accordance with the number of terms that form it (3), thus:

$$D = ((d1, n1), (d2, n2), \ldots (dN, nN)) \quad (3)$$

On the other hand, in the same presentation approach, the value of Q is defined as the set of pairs of values (qi, mi) in which qi is the term and mi is the number of specified terms in the question (4). The value of K indicates the number of terms that are different from the query, thus:

$$Q = ((q1, m1), (q2, m2), \ldots (qK, mK)) \quad (4)$$

The similarity measure between Q and D is calculated, among other methods, according to:

- The number of terms that occur both in the request and in the document.
- The number of occurrences in both (query and document) of the specified common terms occurrences.
- Discriminant value or weight xi of term within the collection of documents. This weight xi of one term t1 is determined in accordance with the number of documents in which the indicated term appears.

Thus, the similarity measure is calculated in accordance with the equation (5):

$$sim(D, Q) = Y \forall i \in Q \wedge D \ (ti, ni, mi, xi, N) \quad (5)$$

where: Y is a method for determining the similarity value between the document and the query according to the parameters.

In other methods that use text passages as a processing unit, calculating the similarity between text passages there is the same approach, but the occurrences of terms are replaced with passages in order to measure the similarity between the query and the document in accordance with the similarity of all textual passages. In addition, many of them do not have a clear methodology of segmenting the document into text passages and calculating the similarity measure between documents.

The study [10] shows that in order to measure the degree of text semantic similarity between these texts, they should not be presented in terms expressed in the same words, but in terms expressed in different words. Then to find similarities in this type of representation automatic translations can help for which the PanLex tool is used, which allows the creation of statistical dictionary. If the translation is possible, it means that the term is equivalent to the term in the text, expressed in other words.

Another way to approach this task is to consider it as a Question Answering problem, where one of the texts is a question, and another is the answer. This is the essence of [11], where a model is proposed that measures the degree of similarity in a function if the answer really answers the question.

In most tasks, at the time of text processing, a type of text comparison is performed, in which words are compared with other words, and / or sentences with other sentences.

Basic criteria of proximity comparison

There are methods that count the frequency of words occurrence in the text, comparing with the original text (query) (6).

$$F_{base} = \frac{p}{q} \quad (6)$$

where p is the number of matching words in the query and the text fragment, q is the number of words in the query. It is believed that two words are the same if their initial forms coincide.

Semantic methods

They compare the sentences and not only calculate the words frequency in the text, comparing with the original text (query), and also consider the relationship between the phrases involved into the comparison. For example, the semantic criterion of comparison for proximity (7):

$$F_{\text{semantic}} = \frac{m}{n} \quad (7)$$

where m is the number of matching elements of the meaning in the query and the text fragment, n is the total number of meaning elements in the query.

In general, the approaches described above have characteristics that allow us to distinguish three groups. The first of them considers the frequency of occurrence of n-gram symbols, words and some lexical relations, such as synonyms and hyperonyms. In addition, many of these approaches emphasize the representation of the natural language, then use the similarity algorithms between strings, such as the Jaccard similarity coefficient, which calculates the number of unique terms shared between the two texts; cosine similarity, which measures the angle between the vectors of both collections of words in the text; the Levenshtein distance, which consists of the minimum number of necessary operations to transform one string of characteristics into another.

Textual semantic similarity is intended to capture the point where the meaning of two texts is similar. This concept is broader than to find the degree of textual similarity, as in the case of the above mentioned algorithms, they measure only the number of lexical components that separate both texts, that is, which does not measure the similarity of two texts with respect to the value that should be expressed.

The second group of characteristics is examined in the same way in researches [12], [13], these are measures of words similarity offered by the NLTK tool in the Python programming language.

In this case, the semantic similarity between the two texts is defined as the maximum value obtained between pairs of words.

The third group considers measures based on Corpus, using the indicators offered to text semantic similarity [14].

Thus, the task is to provide automatic detection of the semantic similarity between two compared texts by creating a method that takes into account both the morphological structure of the text and its lexico-semantic content.

III. THE PROPOSED METHOD OF TEXTS SEMANTIC COMPARISON

To achieve this task, the following steps were accomplished:

- Integration of components or steps framed in general principles of the scheme / model of word processing systems is realized, which is necessary for the development of text comparison model.
- The method for segmenting texts in natural language has been developed, which guarantees the extraction of significant text fragments that preserve the meaning of the text.
- The method for automatically comparing two texts in a natural language has been developed, which reveals a semantic similarity, regardless of the words used.

- Based on the developed model of the proposed comparison, algorithms and methods for segmentation and comparison have been developed, which also make it possible to evaluate the similarity of texts of scientific and technical style according to the criteria of correctness and depth.
- Experimental studies of the methods of segmentation and comparison of texts in natural language were carried out.

It should be noted that at the heart of the proposed method there are fragments of texts, that are already text passages with semantic content [4], but not any fragments.

If we take into account that text segment is a certain set of letters, words or phrases that are part of the text, or any segment of speech characterized by semantic independence and obtained as a result of text segmentation, then a significant text passage can be defined as a separate part of the text that has some kind of integrity. A meaningful text passage is obtained through the identification of semantic aspects, so that a meaningful text passage has a complete semantic meaning. At the moment, the application of the concept of meaning and semantic constructions are the basis for developments in the field of the design of information systems and their components such as databases, user interfaces and information systems development process in general [24], [18], [19].

On the other hand, the meaning is the inner content, the signification of something. Therefore, we formulate the definition of a significant text passage as follows: it is text segment without anaphoric references associated with the words of another segment and in which there is at least one verb, the type and category of which expresses the action.

The proposed method allows you to extract meaningful text passages from texts based on retrieval of text passages with changing the stop criterion. Text passages are extracted from the original and compared texts. This process is performed using the method of extracting random passages [2]. The distinction of the proposed method is that the basis for distinguish text passages is the detection of anaphora. Thus, the criterion for stopping the text division into passages will be the absence of anaphoric references outside the segment. This ensures that the extracted text passage will carry a complete semantic meaning.

At the level of the presentation of text passages in semantic diagrams, the number of n-schemes of the passages of the original text and the number of m-schemes of the compared text passages are obtained, which will then be compared in the ratio n: m, but the coincidences will be counted in the total number of n, regardless of the number of schemes of the compared text, in such a way that if one scheme has coincidences with more than one scheme of another text, this will be considered as the main factor of coincidence. Fig. 1 shows the semantic schema of text passage. To compare two passages represented with such schemas we should compare parameters *r* with coinciding indexes with each other, and then use the equation (9).

At the input of the text comparison process, there are two documents intended for comparison, one of which is the original. At the first analysis level, text passages are extracted, as described in [2], the output of this first level will be a list of meaningful passages from each document that serve as inputs for the next level for the anaphora resolution [3] and the latter, in its queue, arrives at the level of semantic representation of schemes [4]. This representation schema is the input for detecting the level of semantic similarity between textual passages.

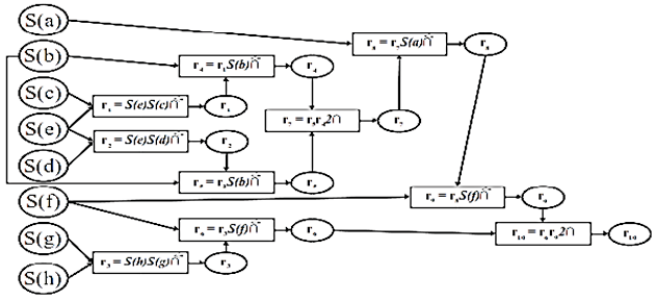

Fig. 1. Example of semantic schema of text passage

At this level, text passages from documents are used as input data and based on their comparison results, the similarity between the two documents is calculated.

When comparing one text with another, each text passage of the compared text should be compared with all passages of the original text in the ratio n: m; from which text passages with great similarity will be selected (8).

$$sim(P1, P2) = max\,\forall ij \in P1 \wedge P2\ sim(P1i, P2j) \qquad (8)$$

The distinction of the proposed criterion for the semantic closeness of textual passages of the original and compared texts is the calculation of the proportion of coinciding elements of meaning, in accordance with the semantic class of words participating in the comparison.

$$F_{semantic/class} = \frac{\sum_i^k \frac{\sum p_j}{l}}{n} \qquad (9)$$

where p is the coincidence factor between the words participating in the comparison for each meaning element, according to the semantic class in the interval [0,1], p = 1 if the word is identical, p = 0 if the word is outside the semantic class and p = (0, 1) depending on the degree of synonymy; l is the number of words for each element of the meaning; k is the number of meaning elements in the text passage of the compared text, n is the total number of sense elements in the text passage of the reference. It is necessary for the expert to determine the degree of synonymy of each semantic class. This can be done by predetermination in the original text.

The definition of correctness and depth depends directly on the goals and objectives of the comparison, and, consequently, the evaluation. A viable criterion of correctness follows from the results obtained in the previous stage, but with respect to the whole text, that is, the similarity coefficient between the standard and the compared text - C, is determined by the formula (10).

$$C = \frac{\sum_1^q F_i}{m}, \qquad (10)$$

where Fi is the result obtained for each i-th comparison; q - the number of text passages of the compared text; and m is the total number of text passages of the original text.

The depth of comparison can be determined in the proportionality of the number of textual passages of the compared text, in relation to the number of text passages of the original, that is, the depth coefficient *S*, is determined by the formula (11).

$$S=\frac{q}{m}, \quad (11)$$

While the score can be denoted by the arithmetic mean of the two previously obtained values of *C* and *S*, the final similarity score between the *R* documents can be determined by the formula (12).

$$R=\frac{C+S}{2} \quad (12)$$

The proposed method of semantic comparison between semantic diagrams of text passages allows you to compare two texts that convey the same or opposite meaning, which are written using different vocabulary, excluding coincidences in similar text passages, unlike existing methods that only measure the number of lexical components contained in both texts or the maximum value of similarity in pairs of words.

IV. Analysis and Comparison of Text Segmentation Based on Meaningful Text Passages (The Results of the Experiment)

We compared the known methods with the proposed method. In particular, the methods of extracting arbitrary text passages and "TextTiling" are compared. In the experiment, for the extraction of arbitrary text passages, the stopping point was used as a criterion.

A hundred original (authentic) texts containing the introduction of scientific papers dedicated to the information systems development extracted from the scientific journal "Izvestia of Southern Federal University" were selected for the experiment; one page each; all texts have been processed, as indicated below.

For each text, ten people produced manual segmentation based on human judgment. To this end, 100 students of 3 and 4 years of education completed the texts processing, with each student producing 10 segmentations. For each text, the limits were chosen as valid, where there were at least 6 coincidences; for example, 10 limits were chosen for the first text: 16, 31, 43, 57, 72, 110, 122, 139, 164 and 190 (Fig. 2).

Then the metric values of "WindowDif" for segmentation were calculated using three algorithms for all 100 texts. In the WindowDif metric [15], a sliding window of length k is used to go through all the text and find the discrepancy between the standard (source text) and the segmentation resulting from the algorithm. Where k is half the average size, which has segments as a part of segmentation. In each window position, the number of constraints existing in the window for both segments is determined, and if the number of boundaries is not the same, the algorithm is violated. Subsequently, all the penalties found in the full text are summed up and this value is normalized in such a way that the metric takes a value from 0 to 1. "WindowDif" takes the value 0 if the algorithm correctly sets all limits and value 1 if they differ from the reference segmentation in all positions of the window.

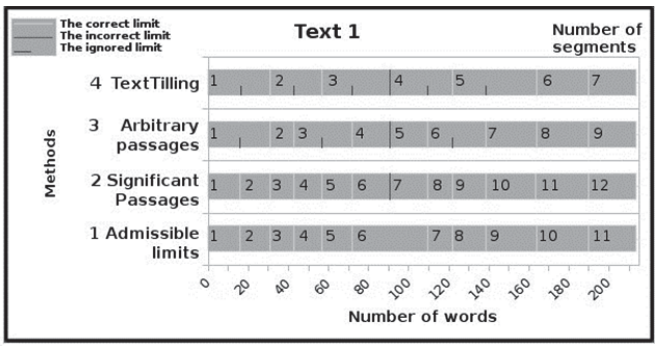

Fig. 2. Comparison of the methods considered for text 1

The results obtained with the help of manual segmentation are compared with the known algorithms – extraction of arbitrary text passages, "TextTiling" and the proposed method. The results of the metric values of WindowDif were converted to percentages of proximity, the average values are shown in Fig. 3; while the performance of three algorithms for 100 texts, ordering the result of each algorithm independently, is shown in Fig. 4.

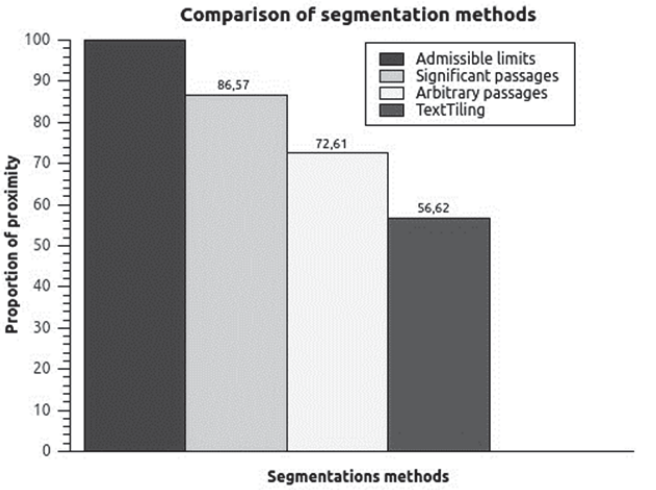

Fig. 3. Average values of the methods considered

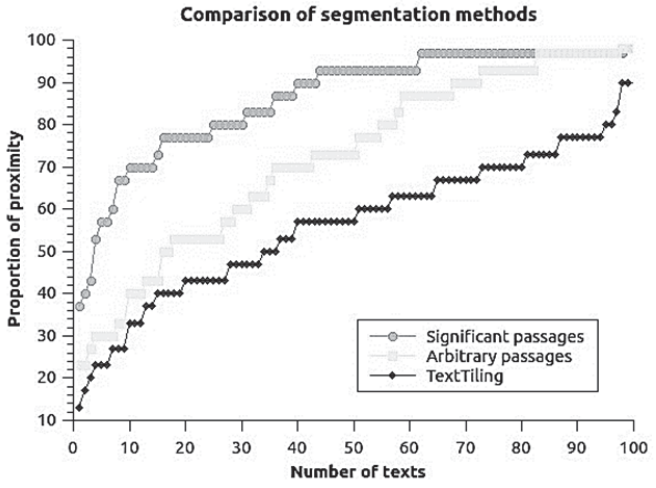

Fig. 4. The performance of algorithms for 100 texts

The results of algorithms sometimes coincide with the limit values indicated as valid for hundred texts. In case of the proposed method, the limits are closest to the sample ones.

V. Analysis and Comparison of the Defining the Texts Similarity Based on Semantic Classes (the Results of the Experiment)

We present the results of the experiment of using the existing methods, the proposed method and the analysis made by the experts. In particular, the following methods and programs are compared:

1) Methods of text comparison based on the Jaccard's similarity index, the cosine similarity and the Levenshtein distance using an online program of similarity algorithms between text strings based on the php programming language [16].

2) The method of latent-semantic analysis and other methods of information retrieval, using for this purpose the online program "plagiarisma.net"; which is based on the use of search engines "Google", "Babylon" and "Yahoo".

3) The program for detecting plagiarism of SFedU called Anti-plagiarism, which is supposedly based on the method of searching and analysis of hidden semantics and on other own algorithms belonging to software developers.

4) The method of determining the similarity for the information search, which is described in the work [17], which is called "F semantic".

For the experiment, four hundred (400) texts were selected, that is: 1) one hundred original (authentic) texts containing the introduction of scientific papers from the collection mentioned in the previous experiment; 2) one hundred texts containing plagiarism (modified text), obtained from the source texts by replacing in the original text of some words with synonyms and phrases; 3) one hundred texts, opposite to the original, which were written intentionally, by replacing some words in the original text with antonyms and phrases of the opposite meaning; 4) one hundred texts of interpretations from original texts written intentionally in response to a question about the general content of the text.

For similarity algorithms between text strings, three hundred texts (of three types) were compared to hundred of original texts, including a comparison with the text itself for the evaluation control, so results were obtained in the form of a percentage similarity between these texts.

For each text, for the plagiarism detection system called "Anti-plagiarism", the original text was at first uploaded in order to make sure that the specified systems have the original text among their databases, then the three remaining texts were given; these systems give the percentage of originality of the uploaded text in relation to the coincidence of their segments with other existing ones. Thus, if the percentage of originality is high, the similarity with the original text is low and vice versa.

For method [17] and the method proposed in the papers [2], [4], text comparisons were made with the original text and with other texts, the result is a similarity expressed in percentages.

For the proposed method [2], [4], a consultation was held with one hundred students from the field of information technology, who were given words and phrases from the original text along with a list of five possible synonyms and no more than two antonyms or phrases with the opposite meaning. Students were asked to assign a degree of similarity of the words on a scale of 1 to 10. The words belong to the same semantic class that was selected from WordNet for the Russian language. For antonyms or phrases with the opposite meaning it was suggested to express their decision, about 60% were recognized. The intermediate results obtained for each word of the semantic class were considered to be a degree of similarity.

Four hundred (400) respondents analyzed four texts each; they were told that the text number one is an original (source) text for comparison with the other three. They were asked to thoroughly study each text in order to get answers to questions about similarity and plagiarism, all in relation to the meaning expressed in the text. This task was carried out using the technological platform of the website: https://toloka.yandex.ru/.

Variants of responses were presented in the Likert quantitative scale. Quantitative results were converted to qualitative in a percentage scale, to compare them with the results of the analyzed methods, taking as a sample the results of the experts' analysis. The above results and their comparison with the existing methods and the proposed method are presented and analyzed in Fig.5.

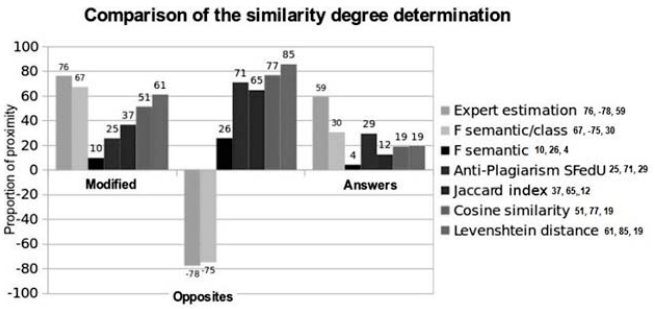

Fig. 5. Average results of similarity calculation

As for the similarity level, on average 91% indicated that one hundred texts of plagiarism are similar or very similar to the originals. 83% indicated that one hundred texts were significantly opposite or completely opposite to the originals. While 75% confirmed that one hundred responses were similar or similar in a small degree; which translates into percentages of similarity in this way: modified texts = 76%; Opposite texts = 78%, and answers to the question about the meaning of texts = 59%.

The results of the text comparison showed (Fig. 6) that the proposed method improves the stability of plagiarism recognition regardless of the percentage of word substitution and improves the detection of similarity by 40% compared to existing systems when replacing more than 50% of the words in the original text.

Of special mention are the results obtained and presented for the Levenshtein distance algorithm, which has a special feature, if the texts are introduced with some changes in the order of the paragraphs in relation to the fragments, the results are significantly reduced. While other algorithms and methods retain the same percentage.

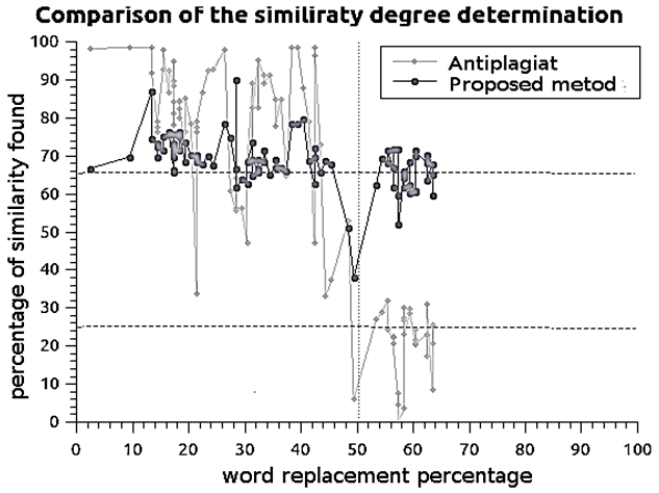

Fig. 6. The performance of methods for 100 texts

This is due to the fact that the Levenshtein distance algorithm is the minimum number of operations required to transform one string of characteristics to another, and, with a change in the paragraphs order, the number of operations increases. But changing the paragraphs order of one text does not change its meaning, and the more cannot disguise plagiarism, in this connection, this algorithm is ineffective for comparative purposes.

VII. Conclusion

A significant scientific novelty in this research is the usage of the term of meaningful text passage in which segments that are compared or processed for any purpose will contain a complete semantic meaning. In addition, after comparing the text segments, the depth and completeness criteria for calculating the similarity are applied. The proposed approach and the method of texts semantic comparison allow to evaluate texts written in natural language and determine the degree of their similarity to the original texts, regardless of the words and syntax used. This is the main difference with respect to existing methods based on the precise identification of words and / or phrases.

The method developed and presented in the paper was tested on texts of scientific and technical style. But, despite the fact that such texts are well suited for automatic processing, not all tasks of natural language processing for this type of texts are completely solved. For example, in information systems like "anti-plagiarism" documents are compared to determine whether one of them was written exactly like the other, but they do not define any coincidence if the plagiarist expounds the author's ideas using other words, synonyms or paraphrases. The proposed in the paper method of texts semantic comparison completely and effectively solves this problem and allows to identify plagiarism at the level of the idea embedded in the text. Moreover, the method makes it possible to determine plagiarism at the level of the opposite meaning.

Experimental studies of segmentation and comparison methods have shown that the proposed method of extracting meaningful text passages allows you to break the text into a number of passages that convey meaning using anaphoric references as a stopping criterion without having to delete words that can convey meaning. This is the main difference from existing methods that eliminate words (stop words) and base their judgments on punctuation marks or statistical criteria that can not guarantee that the text passage will have full semantic meaning.

In addition, the proposed method of semantic comparison between semantic diagrams of text passages allows you to compare two texts that convey the same or opposite meaning, which are written using different vocabulary, excluding coincidences in similar text passages, unlike existing methods that only measure the number of lexical components, contained in both texts or the maximum value of similarity in pairs of words.

The proposed approach has a direct application in systems of plagiarism automatic detection, in order to increase the effectiveness of such detection; as well as in distance education, in order to improve methods of responses assessing. New studies based on the presented approach can contribute to the development of new and innovative methods and systems for the efficiency improving of automatic text processing in natural language, in particular when comparing semantically similar texts written using another dictionary.

Acknowledgment

The reported study was funded by RFBR according to the research project № 17-07-00096 A.

References

[1] Rogozov Y., Sviridov A., "The Concept Of Methodological Information Systems Development", *8th Ieee International Conference On Application Of Information And Communication Technologies, Aict 2014 - Conference Proceedings, 8.* 2014, DOI: 10.1109/ICAICT.2014.7035925

[2] J.G. Bermúdez S., "About the method of extracting meaningful text passages as a base for text comparison", *Journal «Informatization and communication» 3-2016*, p. 147-153. Moscow. ISSN: 2078-8320.

[3] Salguero L. F., "Resolución abductiva de anáforas pronominales", Web:– http://www.http://personal.us.es/fsoler/papers/ivjornadas. Pdf

[4] J.G. Bermúdez S., "Approach to create models of semantic comparison of texts", *Journal «Informatization and communication» 2-2016*, p. 121-126. Moscow. ISSN: 2078-8320.

[5] H. Maurer, F. Kappe, B. Zaka, "Plagiarism - A Survey*", Journal of Universal Computer Science,* 2006, Nº 12, pp. 1050-1084.

[6] L. Chi-Hong, C. Yuen-Yan, "A Natural Language Processing Approach to Automatic Plagiarism Detection", *En Proceedings of the 8th ACM Conference on Information Technology Education (SIGITE'07)*, Florida, EUA: ACM Press. 2007, pp. 213–218. ISBN: 978-1-59593-920-3.

[7] J-P. Bao, J-Y. Shen, X-D. Liu, H-Y. Liu, X-D. Zhang, "Semantic Sequence Kin: A Method of Document Copy Detection", *Advances In Knowledge Discovery and Data Mining*, vol. 3056, Sydney, Australia, 2004, pp. 529-538.

[8] R. Michael, F. Anette, "Automatically identifying implicit arguments to improve argument linking and coherence modeling", *2nd Joint*

*Conference on Lexical and Computational Semantics (*SEM-2013),* Georgia, USA, 2013, pp. 321–333.

[9] E. Agirre, D. Cer, M. Diab, A. Gonzalez-Agirre Weiwei Guo, "Semantic textual similarity", *2nd Joint Conference on Lexical and Computational Semantics (*SEM-2013),* Georgia, USA, 2013, pp. 32–43.

[10] B. Salehi, P. Cook, "Predicting the compositionality of multiword expressions using translations in multiple languages", *Second Joint Conference on Lexical and Computational Semantics (*Sem-2013)*, Atlanta, Georgia, USA. 2013, pp. 134 — 142.

[11] A. Palmer, A. Horbach, M.Pinkal, "Using the text to evaluate short answers for reading comprehension exercises", *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2 (SemEval 2013),* Atlanta, Georgia, USA, 2013, pp. 520–524.

[12] C. Leacock, M. Chodorow, "Combining local context and wordnet similarity for word sense identification", *MIT Press*. 1998. pp. 265–283.

[13] Wu Zhibiao, M. Stone Palmer, "Verb semantics and lexical selection", *Morgan Kaufmann Publishers /ACL,* 1994. pp. 133–138..

[14] R. Mihalcea, C. Corley, C. Strapparava, "Corpus-based and knowledge-based measures of text semantic similarity" *In Proceedings of the 21st National Conference on Artificial Intelligence*, 2006, pp. 775–780.

[15] L. Pevzner, M. Hearst, "A Critique and Improvement of an Evaluation Metric for Text Segmentation", *Computational Linguistics*, 2002, Web: http://people.ischool.berkeley.edu/~hearst/ papers/pevzner-01.pdf.

[16] L. Francesc C., "Algoritmos de similitud entre cadenas de texto (php)", 2015, Web: francescllorens.eu/00tokenizer/dst.php.

[17] R.Y. Vishnyakov, "Development and research of formalized representations and semantic schemes of proposals of texts of scientific and technical style for increasing the effectiveness of information retrieval", Thesis for the degree of Candidate of Technical Sciences. Southern Federal University, 2012, Taganrog.

[18] S. Belikova, Y. Rogozov, E. Borisova, "Semantic approach to information system user interface design", *International Multidisciplinary Scientific GeoConference Surveying Geology and Mining Ecology Management, SGEM 17(21),* 2017, pp. 673-680.

[19] A.N. Belikov, Y.I. Rogozov, O.V. Shevchenko, "Synthesis of the life cycle stages of information systems development", *Advances in Intelligent Systems and Computing*, Volume 763, 2019, pp. 331-337.

[20] M. Kaszkiel. J. Zobel, "Effective Ranking with Arbitrary Passages", *Journal of the American Society for Information Science (JASIS)*, 2001. Web: https://pdfs.semanticscholar.org/64fc/fa996acd5f0c5540a161c359fc343601cdac.pdf.

[21] M. A. Hearst, "TextTiling: segmentating text into multi-paragraph subtopic passages", *Computational Linguistics,* 1997 Web: http://dl.acm.org/citation.cfm?id=972687.

[22] R. Abella-Pérez, J. Medina-Pagola, "An incremental text segmentation by clustering cohesion", in: *Proceedings of the International Workshop on Handling Concept Drift in Adaptive Information Systems: Importance, Challenges and Solutions* (HaCDAIS 2010), 2010, pp. 65–72.

[23] O. Heinonen, "Optimal Multi-Paragraph Text Segmentation by Dynamic Programming". Helsinki: University of Helsinki, 1998. Web: http://www.aclweb.org/anthology/C98-2239.

[24] S. Kucherov, A. Sviridov, S. Belousova, "The formal model of structure-independent databases", *DATA 2014 - Proceedings of 3rd International Conference on Data Management Technologies and Applications*, 2014, pp. 146-152.