UDC 519.766.48 **doi: 10.32620/reks.2023.2.03**

Ihor SHEVCHENKO, Pavlo ANDREEV, Maiia DERNOVA, Olena PODDUBEI

Kremenchuk Mykhailo Ostrohradskyi National University, Kremenchuk, Ukraine

PARAGRAPH-ORIENTED METHODS FOR DETERMINING THE COHERENCE AND THEMATIC UNITY OF SCIENTIFIC AND TECHNICAL TEXTS

The subject of the article *is to determine the degree of scientific and technical text connectedness using statistical calculations.* ***The aim of the scientific investigation*** *is to study the possibilities of using the coherence of fluctuations in the relative frequencies of keywords in paragraphs to determine the lexical coherence and thematic unity of scientific and technical texts.* ***The task*** *is to develop a method for determining the thematic unity of a text at the set of paragraphs level; to develop a method for determining the coherence of a text at the set of paragraphs level; and to test the developed methods on a collection of documents.* ***The methods*** *used are statistical analysis and computational experiment methods. The following* ***results*** *were obtained. The study has shown that it is advisable to cluster paragraphs as points in the keyword space to determine the degree of scientific and technical text coherence at the level of paragraphs. This opens up the possibility of calculating the degree of thematic unity within the clusters and in the entire text. The degree of text fragments and the whole text coherence is determined by analyzing the sequence of paragraph numbers in the clusters. This makes it possible to formally determine the quality of the material presented in a scientific and technical article or in a textbook.* ***Conclusions.*** *The scientific novelty of the study is as follows: there was refined on the method for determination of the connectedness degree (coherence and thematic unity) of scientific and technical texts at the level of paragraphs by implementation of paragraphs clustering in the keywords space, using the calculation of thematic unity degree inside the clusters and in the overall text, as well as through analysis of paragraphs numbers sequence in clusters in order to determine the degree of text fragments and the overall text coherence. The methods are language-independent, based on clear hypotheses, and complement each other. The methods have an adjusting element that can be used to adapt it to different thematic and stylistic areas. It has been experimentally proved that the proposed methods for the determination of scientific and technical text connectedness are efficient and can provide the framework for information technology of content analysis of scientific and technical texts. The proposed methods do not use WEB resources for syntactic and semantic analysis, providing the possibility to use them autonomously.*

Keywords: *text coherence; thematic unity; paragraphs; keywords; relative frequencies; clusters.*

Introduction

Automated text processing is one of the main areas of information processing. An effective means of machine-based extraction of text content is content analysis, which is a methodology of quantitative and qualitative analysis of texts and text arrays for further meaningful interpretation of revealed numerical regularities. Automated text analysis is used in various linguistic tasks (machine translation, machine text recognition, and data extraction). The tasks and problems related to text information processing are in certain directions in the field of natural language processing (NLP). Tasks of this type include keyword search and text coherence assessment. The task of automated keyword search can be solved with many proposed solutions. Most of them use syntactic patterns, which makes computer processing difficult. We have proposed our own approach to solving this problem [1]. There is also the problem of automated text coherence evaluation. In this paper, we elaborate our approach and focus on this task.

In all linguistic dictionaries used by linguists, the term "connectedness" is considered to be one of the principal ones. This concept is the main, inherent feature of the text [1]. We can say that the connectedness of a text is the semantic closeness of the phrases it consists of. The category of connectedness is qualified by researchers as the semantic and structural unity of textual components, which at the semantic level is represented by the subcategory of coherence – internal semantic connectedness between text units; and at the structural level – by the subcategory of cohesion – external connectedness between text units, formally represented by language. The category of connectedness also interacts with the category of integrity, but, unlike the former, the latter is a psycholinguistic category and can be established not in the process of reading, as connectedness, but only after perceiving all the components of the text. It is customary to speak of coherence as a property of the text as a whole, and of cohesion as a type of connection between text elements. It can be argued that cohesion is a set of means of ensuring text coherence at the syntactic and stylistic levels, which also determine the logical and semantic cohesion of sentences [1].

Thus, we emphasize that coherence in the text ensures the integrity and completeness of the utterance, i.e., its connectedness, and that cohesion is only the form of communication through which this integrity is realized. Therefore, in the following, we will pay attention to coherence as a property that needs to be measured when determining the degree of text cohesion.

Our study concerned scientific and educational texts. Such a text should be informative and have a thematic (substantive) unity, i.e., a certain main idea. On the one hand, this facilitates the search for keywords, and on the other hand, it forces the author to make the text coherent. It is coherence that ensures that a scientific and technical text fulfills its informative function. Thanks to coherence, semantic gaps in the text's content are filled in and, as a result, its informative value is increased. In addition, scientific and technical texts are subject to careful control by editors. Therefore, any information support for the process of checking such texts is relevant.

In our study, we use statistical methods to determine the degree of thematic unity and coherence.

1. Analysis of works related and objectives

The NLP methods provide a wide variety of possibilities for data extraction from natural language texts. A good example is paper [2], which describes the application of NLP prognostic analytics aimed at the extraction of certain text entities from twits regarding the evaluation of comments polarity on vaccine quality. This study utilizes the Apache Spark Framework app., which gives the possibility to process significant amounts of data using the distribution method. However, such an approach is impractical in the context of the analysis of a separate document such as an article. On the other hand, it is the analysis of a separate document and its content structure that is considered a relevant objective. Such analysis, as mentioned above, allows not only to classify a document but also to evaluate its quality from the point of view of thematic unity and connectedness.

Such evaluation can be used in various areas of text information processing: search engines and website optimization companies; writing presentation and advertising texts; creation of educational material; and quality control of scientific texts.

The relevance of solving the problem of assessing text connectedness by determining its coherence is confirmed by up-to-date papers that propose methods for assessing the coherence of textual information to solve various problems.

Paper [3] proposes a coherence assessment model based on entity networks that makes it possible to distinguish between connected and unconnected texts by applying a sentence taxonomy based on a semantic representation of entity connectedness. Despite the logical transparency of the model, it produces relatively large discrepancies between the results and expert opinions.

In 2008, a model for assessing text coherence called Entity Grid was proposed [4]. The main idea of this work is to assume that the distribution of key text entities (noun groups present in sentences) follows a certain pattern. The parameter of coherence assessment is the frequency of changes in the role of key entities in the text (subject, direct object, etc.), i.e., the frequency of changes in the accents in the text attracting the reader’s attention is analyzed. In the case of abrupt/even transitions from one key entity to another, the coherence score decreases/increases accordingly. The method requires text parsing. In 2013, a new method for assessing text coherence called Entity Graph was proposed [5]. This paper proposes an unsupervised evaluation of text connectedness based on the construction of a graph in which edges are established between semantically similar sentences represented by vertices. The sentence similarity is calculated based on the cosine similarity of the semantic vectors representing the sentences. This is based on semantic analysis, which in turn is based on syntactic analysis.

Paper [6] discusses some models for evaluating connectedness based on the stirring test. A binary classification is used, in which the model has to distinguish between a document and a reorganized document obtained by randomly shuffling the order of sentences in the document. The models assign a score to each possible position and predict the one with the highest score. One of the limitations of this test is that the accuracy of the model is low, often in the range of 10-20 %. The computational cost is another limitation, often growing linearly with the number of sentences, as it is extremely expensive to evaluate all combinations to find the order with the best score.

In [7], a method of distributed sentence representation using a recurrent neural network is proposed to solve the problem of determining the degree of text coherence. A recurrent neural network was created and trained on a set of Ukrainian-language scientific articles. The result of network training is its ability to perceive and understand a certain text. The universality of this method is questionable. In addition, as in previous works, the use of a neural network is associated with the formation of training corpora of texts. Thus, this way of solving the problem is very difficult.

There are methods that use different types of neural networks to assess text coherence. The corresponding methods are implemented on the basis of convolutional neural networks [8 – 10]. The basic idea is to initially split the text into sentences and then convert it into a sentence-vector or word-vector; this function is performed by the initial layers of neural networks. It is worth noting that this conversion uses pre-trained models of vector representation of sentences or words, which requires the formation of training corpora of texts and regular retraining of the network.

In [11], the authors test whether "skeletons", i.e. key phrases in a text, are a good way to measure text connectedness. Based on the idea of skeleton detection to generate coherent fragments, the authors proposed a new neural network architecture called Sentence/Skeleton NET (SSN) to detect similarities between a pair of sentences or skeletons. The proposed neural model outperforms nonparametric similarity methods such as cosine and Euclidean distance. Of course, such a network also requires training and retraining.

We should also highlight works related to paragraph processing [12, 13]. In [12], sentence parsing is performed by dividing a paragraph into several sentences using a pre-trained Punkt tokenizer for the English language. The Punkt sentence tokenizer is available in the nltk.tokenize module and is provided by the Natural Language Toolkit (NLTK). This tokenizer divides the text into a list of sentences using a model building algorithm for word abbreviations, phrases (combinations of words), and sentence starters. The method requires training on a large collection of plain text before it can be used.

To measure coherence between sentences in a paragraph, [13] proposes a model in which each word is represented as a numerical vector, and for measurement purposes, a sentence is considered the smallest unit of the connected text. To evaluate the integrity of a paragraph, the thematic correlation between sentences is estimated. To evaluate the thematic dependence and cohesion of sentences, sentence matrices are formed using the word2vec vector. Local connectedness is understood as the dependence of sentences in a paragraph. The first step in this process is to estimate the sentence dependency of each paragraph as local cohesion, and in the second step, the paragraph dependency is considered as global cohesion. This means that the next sentence correlates with (n-1) previous sentences according to conditional probabilities. However, the model can assess global coherence only through the thematic unity of sentences and their collections.

The application of deep-learning neural networks for the extraction of hidden information from text data frequencies faces the problem of unbalanced data, which is highlighted in this paper [14]. Such an absence of balance is caused by the necessity to perform preliminary thematic structuring in a large number of texts. The authors propose utilizing the additional neural network as the aforesaid problem solution. But this requires additional training of such a network, which in turn leads to searching for a way to generate structured data on the basis of analysis of certain text corpora. The quality of such texts under their significant quantity needs verification using a particular tool set.

Paper [15] considers an issue arising during text classification using RNN and CNN and their variations. The issue mentioned is that the neural network often focuses not on the global sentence context and the entire text, but rather on the peculiarities of local sentences. It is proposed to solve the problem by applying an additional model of self-attention, which must be focused on key aspects of a text. The model includes a range of additional neural networks, which also require training on the basis of texts with guaranteed quality. Thus, there is a need for simple tools for thematic structuring and text quality verification with regard to its thematic unity.

At the end of the review, we note the works [16] and [17], which present a detailed description of the automatic coherence analysis software application TAACO. The tool widely uses various methods of syntactic analysis to calculate a set of parameters, such as the number and proportion of lemmas with one content, the type-token ratio for bigrams and N-grams, and the role of conjunctions in entity cohesion. The tool calculates average sentence and paragraph matches for all lemma matches, content word matches, and noun, verb, adjective, adverb, and pronoun matches. WEB resources are used to implement all these methods.

The exploration of known ideas has made us sure that none of the authors use the analysis of paragraphs as independent structural units of the text that have thematic unity. It should also be noted that none of the works listed deals with the thematic unity of texts. Meanwhile, this property also significantly affects text connectedness. If a text does not have thematic unity, it is highly unlikely that the whole text will be coherent. However, if a text features the main topic, the concept of thematic unity becomes relevant.

Therefore, in our work, we propose an additional indicator, namely, the indicator of thematic unity. It is desirable that both criteria – coherence and thematic unity – are measured on a single scale. In most of the papers reviewed, the connectivity indicators are measured on-scale [0...1] by reason of the wide utilization of neural networks of various configurations, in which the softmax function is used in the output perceptron layer, which has an output result in the scale [0...1].

In texts containing argumentation and a clear sequence of thought development, which is inherent in professional speech, paragraphs contribute to the creation of the most important feature of a scientific (scientific and professional, etc.) text – the logical sequence of the material. A paragraph of scientific speech is a structural, thematic, and communicative unity. Thematic unity of a paragraph is manifested in the coverage of only one topic [18].

The purpose of this paper, is to develop a technically simple method for measuring the degree of coherence and thematic unity of scientific and technical texts using available statistical methods within the framework of quantitative content analysis. Our approach enables the analysis of a single document without the application of statistical data on the corpus of documents and large dictionaries.

For of our work, the following tasks are set and solved:

– to develop a method for determining the thematic unity of the text at the level of a set of paragraphs;

– to develop a method for determining the coherence of the text at the level of a set of paragraphs;

– to test the developed methods on a set of documents.

2. Method of determining the thematic unity of the text at the level of set of paragraphs

To describe the method for determining the degree of text connectedness, let us devise the following statements:

1. Paragraphs related to the same thematic aspect of an article have a similar set of keywords and distribution of their relative frequencies in paragraphs.

2. In a keyword space, paragraphs related to the same thematic aspect create clusters.

3. The degree of the density of aspect clusters characterizes the thematic unity of corresponding paragraphs. The smaller the mean squared distance between points in the cluster, the greater the thematic unity of paragraphs in this cluster. The concept of thematic unity serves as a local criterion for paragraph information interconnection and directly affects text connectedness.

4. The degree of the proximity of cluster centers to a common center of a keyword space characterizes the thematic unity of a text. The closer clusters' centers to a common center are, the greater is the thematic unity of an entire text.

5. If paragraphs in a created aspect cluster have consecutive numbers, it can be claimed that this text fragment is coherent at the paragraph level.

Based on these statements, let us outline the stages of the method for determining the degree of text thematic unity at the level of paragraphs. It should be clear that this method uses the results obtained in the process of searching for keywords, the general algorithm of which is described above.

Stage 1. Formation of clusters in the keyword space. The number of keywords N (taking into account some keywords in word combinations) determines the dimension of this space. We also know the coordinates of each paragraph point. We use a well-known k-means method. The value of k=N.

Stage 2. Calculation of coordinates of keyword space centroid using relative frequencies of keywords throughout a text.

Stage 3. The iterative procedure of checking clusters for the intersection. For each pair of clusters, the proportion of paragraphs that belong to both clusters is calculated. If there are clusters for which the share of common paragraphs exceeds 0.5, we combine these clusters and calculate the center of a new cluster again.

Stage 4. Calculation of normalized average distance between paragraphs in clusters according to the equation

$$F_1 = \frac{1}{k}\sum_{cl=1}^{k}\frac{1}{2S_{cl}}\sum_{i,\,j\in S_{cl}} d_{ij}^2, \qquad (1)$$

where d_{ij} – is the distance between the i-th and j-th paragraphs in the cl-th cluster;

k is the number of clusters,

S_{cl} is the number of paragraphs in the cl-th cluster. The minimum F_1 is 0. The maximum F_1 is 1.

Stage 5. Calculation of the normalized degree of the proximity of cluster centers to the common center of a keyword space according to the equation

$$F_2 = \frac{1}{2k}\sum_{cl=1}^{k}(\overline{c}-\overline{x}_{cl})^2, \qquad (2)$$

where $\overline{c}$ is a coordinate vector of keyword space center; $\overline{x}_{cl}$ is a coordinate vector of the center of the cl-th cluster. The minimum F_2 is 0. The maximum F_2 is 1.

Stage 6. Calculation of thematic unity of a text.

$$F = 1-[\alpha F_1 + (1-\alpha)F_2], \qquad (3)$$

where α is a weighting factor that balances the influence of local criteria F_1 and F_2. Maximum thematic unity is achieved at $F \rightarrow 1$.

3. Method for determining the degree of text coherence at the level of set of paragraphs

Stage 1. Determination of the degree of coherence of a text fragment corresponding to each cluster. To do this, it is necessary to determine the sequence of paragraph numbers in each cluster and the degree of the irregularity of these sequences.

1.1. Mark with y_{min} – a minimum paragraph number in a cluster, y_{max} – a maximum paragraph number in a cluster.

1.2. Open two arrays – A_1, A_2. The size of the arrays corresponds to the number of paragraphs in a cluster.

1.3. In array A_1, the number of existing paragraphs should be saved and elements should be ranked in ascending order. Array A_2 should be filled with natural numbers ranging from y_{min} to y_{max}.

Stage 2. Calculate the number of discrepancies n_p in arrays A_1 and A_2. Next, the degree of text coherence should be calculated at the level of paragraphs in this cluster with the equation:

$$C_{cl} = 1 - \frac{n_{p_{cl}}}{m_{cl}}, \tag{4}$$

where m_{cl} is the number of paragraphs in the cl-th cluster.

Stage 3. Determine the degree of text coherence at the level of paragraphs with the equation:

$$C = \frac{1}{N_A} \sum_{cl=1}^{N_A} C_{cl}, \tag{5}$$

where N_A is the number of paragraphs in a text.

The maximum degree of coherence is 1.

So, we have two parameters that, in our opinion, characterize the connectedness of a scientific text at the level of paragraphs. This is coherence and thematic unity.

Consequently, a method for determining the degree of connectedness of a scientific and technical text at the level of paragraphs is developed, which is quite interesting because it applies the clustering of paragraphs in a keywords space, calculation of the degree of thematic unity within clusters and an entire text, as well as analysis of paragraph number sequence in clusters to determine the degree of coherence of text fragments and an entire text. This makes it possible to formally determine the quality of the material presented in a scientific and technical article or textbook.

4. Experiments

A set of Ukrainian, Russian, and English-language scientific and technical texts – 20 texts in total – was selected to test the functionality of the methods suggested for text analysis. The set includes scientific and technical articles on various subjects and fragments of study books. The average text volume was about 2200 words. The average number of paragraphs was about 30. The results of the machine analysis for keywords search were compared with the author's sets of keywords in scientific and technical articles. Experts were involved in selecting sets of keywords for fragments of study books.

The results of machine determination of thematic unity and coherence of articles in various scientific fields were compared with the evaluations of expert reviewers. The experts were warned that the ratings of coherence and thematic unity should be on a scale of 0...100 points. This range is easily converted into the range of 0...1, therefore it was taken as a basis. Tables 1 and 2 show the coherence and thematic unity scores for 20 articles. Expert ratings are averaged.

Table 1

The evaluation of coherence

No. of article	Coherence	
	Machine evaluations	Expert evaluations
1	0.43	0.39
2	0.26	0.45
3	0.28	0.36
4	0.35	0.48
5	0.67	0.7
6	0.48	0.53
7	0.78	0.7
8	0.37	0.42
9	0.29	0.36
10	0.64	0.57
11	0.49	0.52
12	0.62	0.54
13	0.27	0.48
14	0.59	0.66
15	0.82	0.75
16	0.45	0.54
17	0.72	0.8
18	0.55	0.63
19	0.44	0.57
20	0.63	0.69
Mean	0.5065	0.557

First, we obtained the experts' scores. Next, the program was tuned by adjusting the value of the coefficient α, which balances the sum of two local criteria that affect the assessment of thematic unity. Since the values of the expert ratings were taken as true, it was necessary to check whether our tool could be adjusted in such a way that the variance of the differences between the machine and expert ratings was minimal (in order to minimize the variance between the machine and the expert ratings). In this regard, a series of experiments, in which the value of the parameter α varied from 0.9 to 0.3 in steps of 0.05, was carried out. It was found that the optimal value equals 0.35. After tuning, the relative discrepancies between the machine scores and the average expert scores were 9.07 % and 7.17 % for coherence and thematic unity, respectively.

Table 2

The evaluation of thematic unity

No of article	Thematic unity	
	Machine evaluations	Expert evaluations
1	0.49	0.52
2	0.31	0.36
3	0.23	0.36
4	0.46	0.5
5	0.52	0.57
6	0.63	0.7
7	0.81	0.75
8	0.46	0.79
9	0.34	0.37
10	0.53	0.58
11	0.42	0.48
12	0.57	0.49
13	0.35	0.45
14	0.62	0.67
15	0.78	0.68
16	0.56	0.58
17	0.64	0.56
18	0.59	0.64
19	0.53	0.57
20	0.78	0.82
Mean	0.531	0.572

5. Discussion

Determining the text coherence at the level of paragraphs is important because the author of a scientific article is qualified enough to arrange sentences within paragraphs in logical order. On the other hand, paragraph arrangement and their overall content quite often require author verification if the author has little experience in article writing.

In particular cases it is expedient to determine the mean square deviation S and to observe which paragraphs and in what number are located within the S, 2S and 3S values from the center of the keyword space after calculating the dispersion of the cluster centers in the keyword space. Potentially, this approach may also be used to evaluate the thematic unity of the text. However, it is necessary to perform additional experiments to determine the adequacy of such thematic unity assessment.

It is also important to note the existence of the problem of unbalanced data in natural language processing technologies based on deep learning networks, in other words, the corpus of texts offered for training or analysis needs thematic structuring for particular subject areas. The methods we propose are well suited for thematic structuring of corpora and selection of texts with the required quality. Furthermore, our methods do not require the use and training of additional neural networks.

The methods proposed for determining the thematic unity and coherence of scientific and technical texts have some advantages over existing methods, namely, they do not require the use of WEB resources aimed at syntactic and semantic analysis, which allows them to act autonomously.

A great number of mathematical expressions in a text lead to certain difficulties in the application of frequency analysis methods. Our approach is not an exception. Therefore, the corpus of texts for testing included articles in which the share of fragments with formulas was no more than 15...20 %. If mathematical expressions make up the significant part of a text, it is impossible to analyze the connectivity using statistical methods. Such texts should be analyzed only by experts in the subject area.

Therefore, at present, machine learning prevails as the main methodology for machine understanding of natural language texts. The application of extremely large amounts of text corpora and very complex deep-learning neural networks allowed us to obtain impressive results. Against this background, traditional methods of quantitative content analysis appear imperfect and ineffective. However, machine learning is based on the same approach as traditional machine analysis, namely the detection of hidden statistical patterns in data sets. On this path, it is possible to obtain positive results by the application of heuristic findings and traditional statistical techniques of analysis. The results obtained may be applied in the future and in machine learning. For example, in work [1], we developed a method of text analysis that differs from the existing methods in that it is based on the identification of positive correlations between the relative frequencies of occurrence of a subset of the most frequent words in paragraphs, which allowed us to identify the keywords and contextual subsets in the texts that feature connectedness at the paragraph level. That is, the hypothesis that there should be certain regularities in the gradual dynamics of keyword occurrence frequencies from one paragraph to another is confirmed if the text under study features connectedness and a certain topic plays the role of a leitmotif.

Conclusions

It has been established that the concept of coherence is an integral feature of the text. The coherence of the text ensures good comprehension and understanding of the text. But connectivity must be provided at different levels. At the level of phrases and sentences, coherence is implemented by syntactic rules. This level of connectivity is called cohesion. At the level of the text as a whole, coherence is manifested in its thematic structure and unified semantics. Therefore, it is generally advisable to define coherence using larger structural units, namely paragraphs. To this level, the concept of coherence is related to the concept of coherence and thematic unity. Coherence is defined as the internal meaningful connection between the text units.

We believe that the criterion of thematic unity is crucial for the assessment of texts devoted to particular aspects of science and technology. Therefore, the work pays attention to the thematic unity assessment. Thematic unity cannot be established during the process of reading, but only after the perception of all components of the text. Both of these text characteristics must be evaluated in many cases.

The existing methods of connectedness evaluation are based either on syntactic rules or on neural network technologies for natural text analysis. Obviously, both approaches require the application of large resources available in the global network. Meanwhile, the coherence and thematic unity of the texts of a particular category, for example, scientific and technical articles, may be assessed by fairly simple statistical methods. That particular direction is under discussion in our paper.

The method for determining the degree of connectedness (coherence and thematic unity) of a scientific and technical text at the level of paragraphs has been improved. It is based on paragraph clustering in keywords space, calculation of the degree of thematic unity within clusters and the entire text, and analysis of paragraph number sequence in clusters to determine the degree of coherence of text fragments and in the whole text. This makes it possible to determine the quality of the material presented in a scientific and technical article or in a manual.

The methods may be used as an auxiliary tool for content analysis of scientific and technical texts.

Further studies are aimed at testing the application of our methods on a wider range of documents, using a set of categorical thesauri to highlight thematic aspects of the text.

Authors' contribution: methods of paragraph-oriented text processing were developed by **Pavlo Andreev**; the software implementation of the methods was created by **Pavlo Andreev**; experimental studies were conducted by **Maya Dernova** and **Olena Poddubei**; review and analysis of the literature, the general formulation of the task and the introductory part of the article were prepared by **Ihor Shevchenko**.

All authors have read and approved the published version of the manuscript.

References

1. Shevchenko, I., Andreev, P., Dernova, M. and Khairova, N. Vykorystannya statystychnoyi modeli kogerentnosti zviaznogo tekstu v iakosti dodatkovogo instrumentu kilkisnogo kontent-analizu [Use of the statistical model of coherence of connected text as an additional tool of quantitative content analysis]. *Visnyk Kremenchuts'koho natsional'noho universytetu imeni Mykhayla Ostrohrads'koho*, Kremenchuk, KrNU Publ., 2021, no. 5, pp. 62-67. DOI: 10.30929/1995-0519.2021.5.62-67.

2. Badi, H., Badi, I., El Moutaouakil, K., Khamjane, A. and Bahri, A. Sentiment analysis and prediction of polarity vaccines based on Twitter data using deep NLP techniques. *Radioelectronic and Computer Systems*, 2022, no. 4(104), pp. 19-29. DOI: 10.32620/reks.2022.4.02.

3. Barzilay, R. and Lapata, M. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 2008, vol. 34, iss. 1, pp. 1-34. DOI: 10.1162/coli.2008.34.1.1.

4. Guinaudeau, C. and Strube, M. Graph-based local coherence modeling. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 2013, vol. 1, pp. 93-103. Available at: https://aclanthology.org/P13-1010.pdf (accessed 12.02.2023).

5. Putra, J. W. G. and Tokunaga, T. Evaluating text coherence based on semantic similarity graph. *Proceedings of TextGraphs-11: the Workshop on Graph-based Methods for Natural Language Processing ACL, 2017,* Association for Computational Linguistics, Vancouver, Canada, August 3, 2017, pp. 76-85. Available at: http://aclanthology.lst.uni-saarland.de/W17-2410.pdf (accessed 03.02.2023).

6. Laban, P., Dai, L., Bandarkar, L. and Hearst, M. A. Can Transformer Models Measure Coherence In Text? Re-Thinking the Shuffle Test. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Short Papers),* Association for Computational Linguistics, August 1-6, 2021, pp. 1058-1064. Available at: https://aclanthology.org/2021.acl-short.134.pdf (accessed 03.02.2023).

7. Pogorilyy, S. D., Kramov, A. A. and Biletskyi, P. V. Metod otsinky kogerentnosti ukrainomovnykh tekstiv

z vykorystanniam zgortkovoi neyronnoi merezhi [Method for coherece evaluation of ukrainian texts using convo-lutional neural network]. *Zbirnyk naukovykh prats' Viys'kovoho instytutu Kyyivs'koho natsional'noho universytetu imeni Tarasa Shevchenka – Collection of Scientific Works of the Military Institute of Kyiv National Taras Shevchenko University*, 2020, vol. 65, pp. 64-71. DOI: 10.17721/2519-481X/2019/65-08.

8. Li, J. and Hovy, E. A model of coherence based on distributed sentence representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, October 25-29, 2014, Doha, Qatar, pp. 2039-2048. DOI: 10.3115/v1/D14-1218.

9. Cui, B., Li, Y., Zhang, Y. and Zhang, Z. Text Coherence Analysis Based on Deep Neural Network. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, November 6-10, 2017, Singapore, Singapore, pp. 2027-2030. DOI: 10.1145/3132847.3133047.

10. Wadud, Md. A. H. and Rakib Md. R. H. Text Coherence Analysis based on Misspelling Oblivious Word Embeddings and Deep Neural Network. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 2021, vol. 12, no. 1, pp. 194-203. DOI: 10.14569/IJACSA.2021.0120124.

11. Xu, J., Ren, X., Zhang, Y., Zeng, Q., Cai, X. and Sun, X. A Skeleton-Based Model for Promoting Coherence Among Sentences in Narrative Story Generation. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, Association for Computational Linguistics, Brussels, Belgium, pp. 4306-4315. DOI: 10.18653/v1/D18-1462.

12. Putri, E. H., Fadilah, D. R., Ivan, Suhartono, D., and Wiannastiti, M. Thematic Development for Measuring Cohesion and Coherence Between Sentences in English Paragraph. *Fourth International Conference on Information and Communication Technologies (ICoICT)*, Bandung, Indonesia, 2016, pp. 54-59. DOI: 10.1109/ICoICT.2016.7571883.

13. Abdolahi, M. and Zahedi, M. A new model for text coherence evaluation using statistical characteristics. *Journal of Electrical and Computer Engineering Innovations*, 2018, vol. 6, iss. 1, pp. 15-24. DOI: 10.22061/JECEI.2018.799.

14. Li, K., Yan, D., Liu, Y. and Zhu, Q. A network-based feature extraction model for imbalanced text data. *Expert Systems with Applications*, 2022, vol. 195, article no. 116600. DOI: 10.1016/j.eswa.2022.116600.

15. Wang, X., Chen, Y., Liu, W. and Tai, W. Research on Text Classification Model Based on Self-Attention Mechanism and Multi-Neural Network. *3rd International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE2022)*, October 21-23, 2022, Guangzhou, China. Available at: https://ceur-ws.org/Vol-3304/paper30.pdf (accessed 30.12.2022).

16. Crossley, S. A., Kyle, K. and Dascalu, M. The Tool for the Automatic Analysis of Cohesion 2.0: Integrating Semantic Similarity and Text Overlap. *Behavioral Research Methods*, 2019, vol. 51, iss. 1, pp. 14-27. DOI: 10.3758/s13428-018-1142-4.

17. Crossley, S. A., Kyle, K. and McNamara, D. S. The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*, 2016, vol. 48, iss. 4, pp. 1227-1237. DOI: 10.3758/s13428-015-0651-7.

18. Le, Elisabeth. The role of paragraphs in the construction of coherence text linguistics and translation studies. *International Review of Applied Linguistics in Language Teaching (IRAL)*, 2004, vol. 42, iss. 3, pp. 259-275. DOI: 10.1515/iral.2004.013.

Received 06.03.2023, Accepted 20.05.2023

АБЗАЦ-ОРІЄНТОВАНІ МЕТОДИ ВИЗНАЧЕННЯ КОГЕРЕНТНОСТІ ТА ТЕМАТИЧНОЇ ЄДНОСТІ НАУКОВО-ТЕХНІЧНИХ ТЕКСТІВ

Ігор Шевченко, Павло Андреєв, Майя Дернова,
Олена Поддубей

Предметом статті є визначення ступеню зв’язності науково-технічних текстів за допомогою статистичних обчислень. **Метою роботи** є дослідження можливостей використання когерентності коливань відносних частот ключових слів в абзацах для визначення лексичної когерентності та тематичної єдності науково-технічних текстів. **Завдання** полягає у розробці методу визначення тематичної єдності тексту на рівні сукупності абзаців; розробці методу визначення когерентності тексту на рівні сукупності абзаців; випробуванні розроблених методів на колекції документів. Використовувані **методи** − це методи статистичного аналізу та методи обчислювального експерименту. Отримані наступні **результати**. У процесі дослідження показано, що для визначення ступеню зв’язності науково-технічного тексту на рівні абзаців, доцільно робити кластеризацію абзаців, як точок у просторі ключових слів. При цьому відкривається можливість обчислення міри тематичної єдності всередині кластерів і усього тексту. Ступінь зв’язності фрагментів тексту та всього тексту в цілому визначається за допомогою аналізу послідовності номерів абзаців у кластерах. Це дає можливість формально визначити якість викладення матеріалу у науково-технічній статті або навчальному посібнику. **Висновки.** Наукова новизна полягає у наступному: ми вдосконалили метод визначення ступеню зв’язності (когерентності та тематичної єдності) науково-технічного тексту на рівні абзаців за рахунок того, що застосовуємо кластеризацію абзаців у просторі ключових слів, обчислення міри тематичної єдності всередині кластерів і усього тексту, а також аналіз послідовності номерів абзаців у кластерах для визначення ступеню когерентності фрагментів тексту та усього тексту в цілому. Методи не залежать від мови, засновані на зрозумілих гіпотезах та вдало доповнюють один одного. Методи мають елемент налаштування, якій можна використовувати для адаптації під різні тематичні та стилістичні напрями. Експериментально доказано, що запропоновані методи визначення зв’язності науково-технічних текстів працездатні та можуть слугувати основою для створення інформаційної технології контент-аналізу науково-технічних текстів. Методи не потрібують використання вебресурсів, спрямованих на синтаксичний та семантичний аналіз, дозволяючи діяти автономно.

Ключові слова: когерентність тексту; тематична єдність; абзаци; ключові слова; відносні частоти; кластери.

Шевченко Ігор Васильович – д-р техн. наук, проф., проф. каф. автоматизації та інформаційних систем, Кременчуцький національний університет імені Михайла Остроградського, Кременчук, Україна.

Андреєв Павло Ігорович – асп. каф. автоматизації та інформаційних систем, Кременчуцький національний університет імені Михайла Остроградського, Кременчук, Україна.

Дернова Майя Григорівна – д-р пед. наук, проф., проф. каф. автоматизації та інформаційних систем, Кременчуцький національний університет імені Михайла Остроградського, Кременчук, Україна.

Поддубей Олена Вікторівна – канд. пед. наук, доц., доц. каф. автоматизації та інформаційних систем, Кременчуцький національний університет імені Михайла Остроградського, Кременчук, Україна.

Ihor Shevchenko – D.Sc., Professor, Professor of Automation and Information Systems Department, Kremenchuk Mykhailo Ostrohradskyi National University, Kremenchuk, Ukraine,
e-mail: ius.shevchenko@gmail.com, ORCID: 0000-0003-3009-8611.

Pavlo Andreev – PhD student of Automation and Information Systems Department, Kremenchuk Mykhailo Ostrohradskyi National University, Kremenchuk, Ukraine,
e-mail: pavel.andreyev.98@gmail.com, ORCID: 0000-0003-4368-9584.

Maiia Dernova – Doctor of Pedagogical Sciences, Professor, Professor of Automation and Information Systems Department, Kremenchuk Mykhailo Ostrohradskyi National University, Kremenchuk, Ukraine,
e-mail: maya.dernova@ukr.net, ORCID: 0000-0003-4545-5247.

Olena Poddubei – C.Sc., Associate Professor, Associate Professor of Automation and Information Systems Department, Kremenchuk Mykhailo Ostrohradskyi National University, Kremenchuk, Ukraine,
e-mail: helenunderoak@gmail.com, ORCID: 0000-0003-1915-4889.