IEEE Second International Conference on Data Stream Mining & Processing
August 21-25, 2018, Lviv, Ukraine

Computer System of Building of the Semantic Model of the Document

O.S. Volkovsky
Department of Computer Science and Information Technologies,
Oles Honchar Dnipro National University
Dnipro, Ukraine
didivave@mail.ru.

Y. R. Kovylin
Department of Computer Science and Information Technologies,
Oles Honchar Dnipro National University
Dnipro, Ukraine
kovilin.yegor@gmail.com.

Abstract **– The analysis of the existing approaches to the development of applied models of natural language was performed. The algorithms of building of semantic representation of the text were developed. The structure of the computer system of building of the semantic program model of the document was examined. The application of the developed system for automatic determination of coherence of the text in the natural language was described.**

Key words — semantic network, inquiry/response system, automated text processing

I. Introduction

The development of applied program systems of automatic texts processing means the selection of one or another way of description and implementation of the model of natural language available to ECM. Due to the fact that the language is rather non-formalized system with instability and non-uniformity of its own rules then the main problem is a difficulty of description of semantic characteristics of the text at the level of algorithmic representation. As long as the natural language is not just a set of words based on some grammar constituents (the top-priority focus of automated texts processing tasks is the obtainment of the very sense bearing text) this, consequently, makes many developers to consider the semantic relations not only between certain words but also between sentences and even between the documents. In most cases the semantics and understanding of the text by the machine mean the following: we cannot speak about the semantics in the event of entering of the text into ECM memory and printing of it with the use of printing device; however, we can speak about the semantic understanding of the text by the computer if the text is processed in one or another way and, as a result of such processing, the reader obtains a new text that is understandable and adequate for him or her (for example, the translation into other language). From this perspective we can speak with confidence about the importance of creation of automated approach to the program simulation of the natural language. The present paper deals with the structure of the system of automatic building of computer model of obtainment of semantic relations in the text that is available for the further applied program implementation of systems of automated text processing.

II. The Analysis Of Existing Computer Models Of Texts

There are several classes of systems of text processing in present-day computer linguistics; they differ against each other both in complexity of data processing and complexity of the intelligent component. Conditionally these systems can be divided into three types according to language models forming their basis: Chomsky generative grammar, semantic network and tools of neuronal networks. Due to the fact that the necessity of applied implementation in this case ranks above the theoretic element, let's consider the certain program systems of every specified class. As a first step let's consider the program of generation of test assignments for extra mural study of students based on the paradigms of Chomsky formal grammar [1]. Generative constituent grammar is based on the axiom of the existence of the phenomenon of linguistic competence consisting in ability of human to master and to understand natural human language. According to this the generative grammar establishes a goal to simulate this ability within generation of correct sentences using certain finite set of rules, alphabet and initial symbol of the sentence – the immediate constituents. Theoretically, plenty of immediate constituents is unrestricted and infinite; in practice the very language, subject field, working text corpus and possibilities of ECM essentially reduce the size of the plenty of immediate constituents.

The technology of semantic networks that is the next step in the development of texts processing industry came into widespread acceptance in the field of automatic texts processing. The semantic network is a graph the points of which comprise the semantic units, and the arcs of which describe the notional relations between them. Typically, the semantic units mean the single word or the sentence or even the entire document. The practical application of the semantic network to the task of text processing is well represented in the paper [2] – the system of automatic counseling. Developers focus on generation of base of knowledge of certain subject field for the provision of dialogue with the user according to corresponding questions. There is provided to use the semantic network for storage of extracted knowledge on the basis of training corpus represented by the sets of pre-readied boilerplate phrases-answers.

Currently the most automated tools applicable to the task of automatic generation of texts are the applied artificial intelligence methods – implementation of automated texts processing with the use of neuronal networks. The artificial neuronal networks are in general use today for the purpose of handling of various applied problems, including the tasks of automatic language processing. For the purpose of assessment of quality of neuronal networks in the task of automatic generation of texts let's consider the paper [7] where the recurrent network is used for drafting the description of products in some on-line store. As we can see,

978-1-5386-2874-4/18/$31.00 ©2018 IEEE

the results are rather mixed. The main advantage of this approach is a complete automation of the process of text generation, high degree of system adaptivity and low costs for its setting and implementation. However, some problems of notional rubbish production are evident. The reason is that in spite of the false availability of the intelligent processing, the system does not see the meaning of described and generated text, basing exclusively on pre-given patterns – teachers.

In the event of building of program system of automatic texts processing with intelligent semantic components the notional relations between the elements of the text are particularly important characteristics of the text, so the semantic networks are the best choice for description of model of natural language on the basis of which the assessment and the relationship extraction is carried out. Therefore, it opens a question of building of alternative ways of necessity of compilation of model base of knowledge on the basis of which the semantic network will be formed. The approach used in the paper [2] forms the closed system, the results of work of which do not exceed the limits of the base of knowledge added to it, whereas the most important parameters of the task of automatic texts processing is the adaptivity and the generality of applied use of development.

The creation of the semantic net of a text is not a new task. At this time, there are several approaches to the computer processing of the semantic nets for both Slavic and English languages. The basis of all these approaches, which form the basic relations between elements in the text is the ontology production model [3], an example of which is illustrated in formula (1):

$$Qi = Pi(A, B) \tag{1}$$

where i is the product's *Q* name, with which the product stands out from the all working set (as a name can be used a running number from the set of products, which is stored in the system's memory), $Pi(A,B)$ – a predicate of relations, which describes how the element of language *A* is affected by *B*.

For example, the word "burn" can be described as (fire, action). The practical application of this technology is described detailed in the work of [3], on the basis of which it is created a semantic meta-description of the test document for the future semantic search. The meta description is defined as the triplets, which contain the sentences of the original text. The key feature within the frame of our work is that the basic system data is formed on the basis of the previously manually marked body of the Russian language.

The increasing scope of the semantic net led to the creation of the net formation approach as a net model, described in [4].

In this case, the predicative relation is described by formula (2):

$$Q = P(I, Ci...Cn) \tag{2}$$

where *I* – an information units set, $Ci...Cn$ - a set of the links types between information units. Such nets are often used as the documents search models in the body, as it is well suited for the links description of the set of texts against each other. The applied application of the net models within the task of the document's semantic models formation is still in question.

The further development of the semantic nets technology received in work of [5]. The suggested semantic Q-net has a pyramidal structure and, therefore, all text parts, reflecting the essential units of the subject area or integrated complex objects, for detection of which the special relations were introduced, will always be reflected in this net by the corresponding vertices. Each network pyramid defines a certain text fragment of one of four types. Moreover, Q-nets have the properties of homogeneity and hierarchy, allowing the formation of relationships between semantic objects. It is expected in future that by representing with the help of one Q-net the texts selection of this subject area and using the mechanisms for formation of the generalized objects class definitions and relations in the pyramidal nets, it will be possible to automate the process of the ontology construction of this subject area.

An interesting practical development with the use of semantic nets is the forming system of a semantic net from the weakly structured text sources, described in the work of [6]. The authors of the work offer an approach for the automatic recovery of the article's sections structure of the open dictionary Wiktionary. The peculiarity of this approach is the development of a certain rules system, on the basis of which a semantic program model of the article is created.

III. Novelty Of Research And Comparison Of Existing Approaches

Most of the applied developments of the computer systems with the use of the semantic nets suppose the use as the starting knowledge basis some block of texts, which contain a previous linguistic annotation. In such a way, it was described in the work of [3] a system, which was initially based on the articles of the Russian national corpus, which is not only closed for the public use, but also contains the markings solely based on Russian-language materials. An alternative for the automated text processing of the other flexional rich languages, as for example Ukrainian, doesn't exist at this moment. The further improvements of the semantic nets, as in the works of [4] and [5] touched upon a question of modification of the net structure itself, and not of the automation methods for the formation and processing of the original system data and it did not find the applied application within the frame of our task. The alternative approach for the net formation is the use of some rules system, as was described in the work of [6]. Such approach allows avoiding of a previous necessity of the linguistic text annotation. However, the use of such method for the natural language is limited, as due to the lack of enough formalization, high flexion, a large number of exceptions and the properties of language variability, it is not possible at the moment to create and effectively to process such a set of rules at the applied software level.

The main task of the developed within the frames of this work approach to the construction of a semantic net is overcoming of the necessity of the previous receiving of any linguistic knowledge. The described technique allows to build semantic relations between the elements of the document and to put them in line with the numerical semantic weight, forming, in such a way, a program model of the document. Herewith, it is used in the knowledge base system neither linguistic annotation of the document, as in the work of [3-5], nor the system of language rules, as in the work of [6], that allows the receiving of the program semantic models of the documents with the high level of adaptability and independence from language.

IV. The System Of Building Of Semantic Model Of The Document

The algorithm of system's work is represented in the Fig. 1. The first stage after document download and syntactic analysis performance (detection of sentences and words) is a determination of the part of speech for every found word. For this purpose the training sample collection from 15 thousand of words and parts of speech corresponding to them was included into the system. The training of naive Bayes classifier is carried out for every element on the basis of three types of flexions (two and three last letters of the word and flexion obtained through Porter's stemming algorithm), afterwards the establishment of part of speech for every selected word in the text is carried out on the basis of trained model. The final dictionaries of auxiliary parts of speech were included into system for the purpose of accuracy increase.

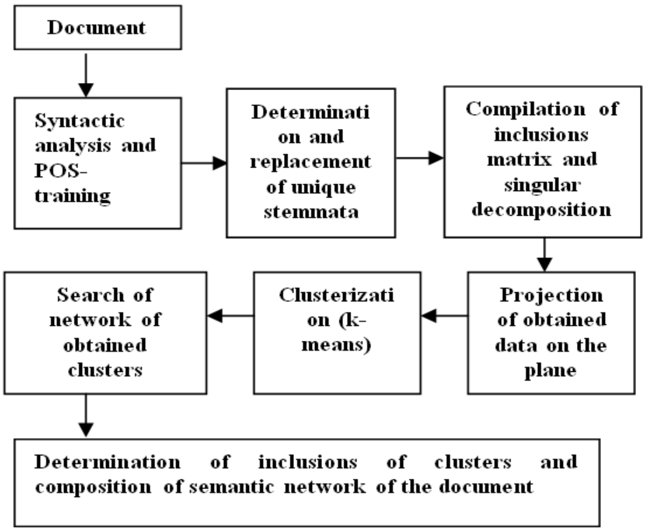
Document
Syntactic analysis and POS-training
Determination and replacement of unique stemmata
Compilation of inclusions matrix and singular decomposition
Projection of obtained data on the plane
Clusterization (k-means)
Search of network of obtained clusters
Determination of inclusions of clusters and composition of semantic network of the document

Fig. 1. General structure of work of system of building of semantic network of the text

The determination of unique stemmata is carried out after this stage: the clipping of prefix and flexion is conducted for every pair of words by Porter algorithm – if the length of the maximum common part exceeds or is equal to Levenshtein distance for this pair of words then the analyzed word is replaced by obtained stemma. As a result the text has the following form (Table I) – the system determined the part of speech and number of inclusions of stemma into text for every stemma. All words marked as auxiliary parts of speech are removed from the text at this stage.

The matrix *N*M* the values of which are determined by quantity of inclusions of stemmata into the sentence is composed for every stemma (with a total quantity of *N*) and every sentence (with a total quantity of *M*). The operation of singular value decomposition and projection of obtained data on the plane is conducted on the obtained matrix.

TABLE I. THE RESULT OF SYNTAX PROCESSING

<table>
<tr><th>Text before being processed</th><th>Text after processing</th></tr>
<tr><td>Today there are various books, video courses on software engineering and other ways to learn to create programs and applications quickly and relatively inexpensively.</td><td>|2||[day{ADV}]|3|[there a{V}]|2|[vari{A}]|2|[book{S}]|3|[videocours{S}]|22|program{S}]|4|[othe{S}]7|[way{S}] |2|[quickl{ADV}]|2|[inexpensive{A}]|3|[to learn{V}]|8|[to crea{V}]|22|[program{S}] |2|[application{S}]</td></tr>
</table>

Due to the fact that the singular value decomposition is stable, we can omit those values of left and right matrix that correspond to low singular values, and to keep only the first two that represent the vectors of coordinates of two-dimensional plane for stemmata and sentences. The obtained projection is represented in the Fig. 2.

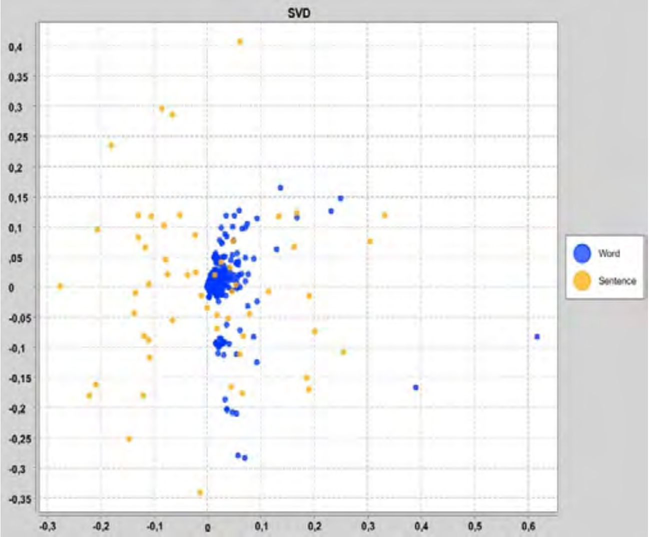

Fig. 2. Projection of singular decomposition: Word – coordinates-stemmata, Sentence – coordinates-sentences of text network

The next step is a clusterization of points-coordinates for stemmata and sentences by *k-means* algorithm. Quantity of clusters for stemmata and sentences *cl* is determined according to the formula (3):

$$cl(W,W_U)=\frac{count(W)}{count(W_U)} \qquad (3)$$

where *W* means the words, W_U means stemmata. Centroids of clusters – stemmata are the coordinates of stemmata with the maximum number of inclusions into the text determined according to the formula (4):

$$Cst(W_U)=\max(W_0...W_{cl}) \qquad (4)$$

where $W_0...W_{cl}$ are the weights of stemmata. Centroids of clusters-sentences are the coordinates of sentences with the maximum total weight of stemmata determined according to the formula (5):

$$Cs(W_S)=\max\left(\sum_{i=0}^{SN}W_i\right) \qquad (5)$$

where W_S is a sentence, W_i is a weight of stemma in the sentence, *SN* is a quantity of stemmata in the sentence. The result of such operations can be seen in the Fig. 3 (for stemmata) and in the Fig. 4 (for sentences).

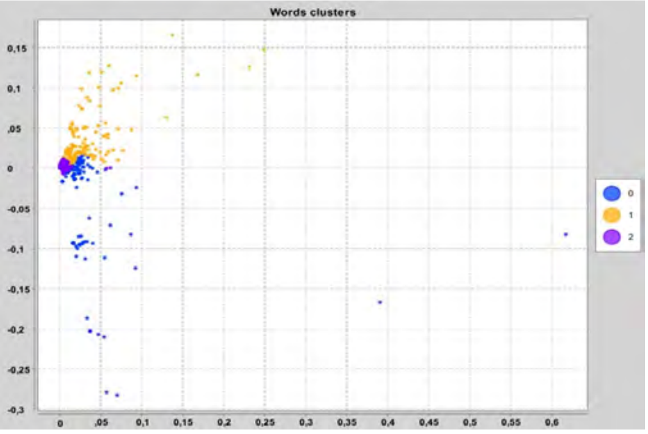

Fig. 3. Projection of clusterization for stemmata. 0, 1, 2 – numbers of clusters

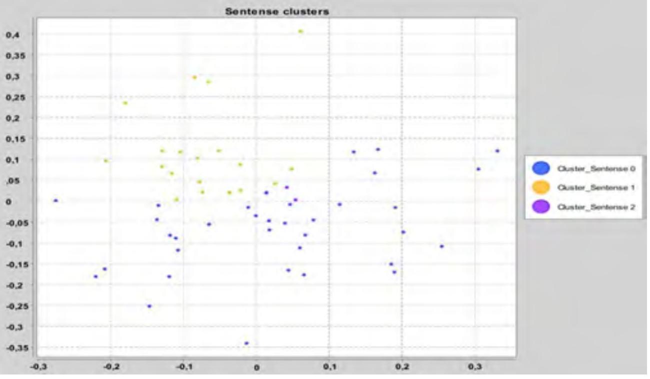

Fig. 4. Projection of clusterization for sentences. Cluster_Sentense 0, Cluster_Sentense 1, Cluster_Sentense 2 – numbers of clusters

The formation of semantic network of the document becomes the final stage. The outline of convex figure is built by Jarvis algorithm on the basis of coordinates of points of every cluster-stemma. The weight – number of stemmata included into it – is determined for every cluster – stemma from whence the semantic graph of relations of clusters in the descending order of their weight is built. Hit of points forming every cluster – sentence is verified for every figure of clusters – stemmata obtained by Jarvis algorithm. If such points are found then the cluster of sentence is connected in the network to the cluster-stemma where the weight of relation is a number of points having hit into the outline of cluster-stemma. The result of system's work is represented in the Fig. 5.

V. Results Of Experiments

Following on from the obtained results the formation of mathematical model of the document in the systems of automated texts processing becomes possible. The system was verified with the use of texts created as a result of automatic generation on the basis of patterns for the purpose of verification of the adequacy of obtained model. Such texts are statically correct, but they have weak notional relations between their parts. The result of processing of such text is represented in the Fig. 6 (projection of clusters – stemmata and sentences) and in the Fig. 7 (resulting semantic network).

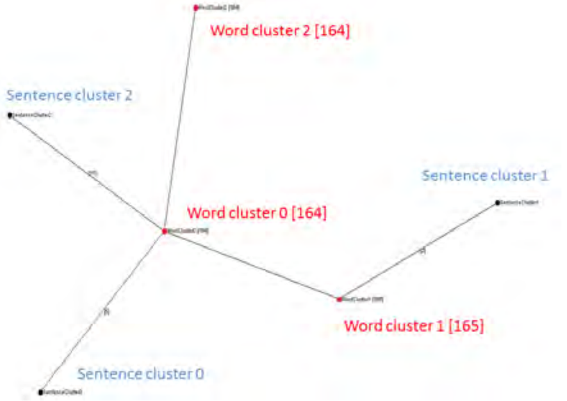

Fig. 5. The result of system's work – semantic network of the document. WordCluster corresponds to clusters – stemmata, SentenceCluster corresponds to clusters – sentences.

In spite of the fact that the size of the automatically generated text was coincident with the example provided before, its semantic network has the apparent differences. This observation enables to make an assumption not only about the adequacy of the semantic model, but also about the possibilities of its application beyond the task of building of inquiry/response systems. Such data as the quantity of clusters – sentences and clusters – stemmata, quantity of relations and their weight, weight of clusters – words, included into the semantic network, can be used for training of models, included, for example, to the systems of automatic determination of plagiarism or text coherence.

Program model of the text obtained in such way can be used for the purpose of building of systems with complex intelligent component of semantic analysis of the text in the natural language. The example of such application is a system of definition of text coherence, described in the paper [8]. Taken data are transferred to the ingress of asynchronous neuronal network that makes a decision on text coherence on the basis of data from model training corpus. It should be considered that described semantic characteristics depend on text size, sob taken data require prenormalization. For this purpose the corpus from eighty texts on the topic of information technologies, astronomy and incoherent texts obtained due to services of frequency autogeneration was compiled; each text was characterized by two values – standard text size W_N obtained according to the formula (6):

$$W_N = \frac{W_i - W_{\min}}{W_{\max} - W_{\min}} \qquad (6)$$

where *Wi* is a total quantity of words, W_{min} and W_{max} is a minimum and maximum quantity of words in the training corpus and normalized semantic value SN, obtained according to the formula (7):

$$S_N = \frac{W_U}{W} \bullet \frac{CW_C}{CW} \qquad (7)$$

where W_U is a total quantity of stemmata, W is a total quantity of words, CW_C is a quantity of clusters – stemmata, related to clusters – sentences, CW is a total quantity of clusters – stemmata. Data obtained in such way compile the training sample collection for neuronal network.

Sample collection from 20 texts, both non-coherent (auto-generated) and real scientific texts on topics of astronomy, information technologies and economics was drawn for the purpose of system testing. In addition to it, the sample collection included also the text assembled from various coherent parts of texts of one topic, but not related semantically in general. The results of texts processing are presented in the Table II, where «n» corresponds to auto-generated text, «z» means connected text, «s» is a text assembled from various parts, 1 – prognosis points to text coherence, 0 – prognosis points to texts incoherence.

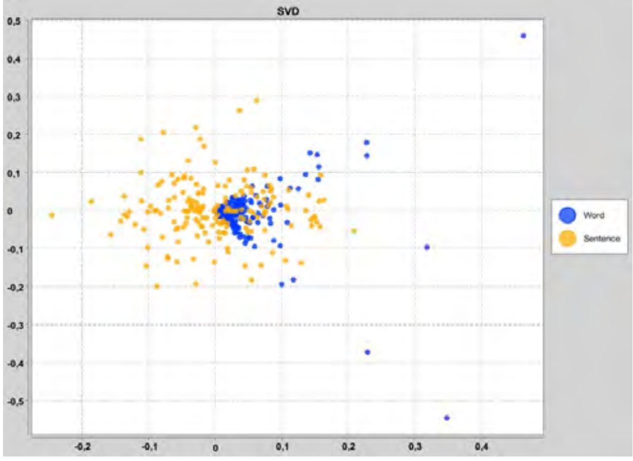

Fig. 6. Projection of singular decomposition for the text with weak semantic relations: Word – coordinates – stemmata, Sentence – coordinates – sentences.

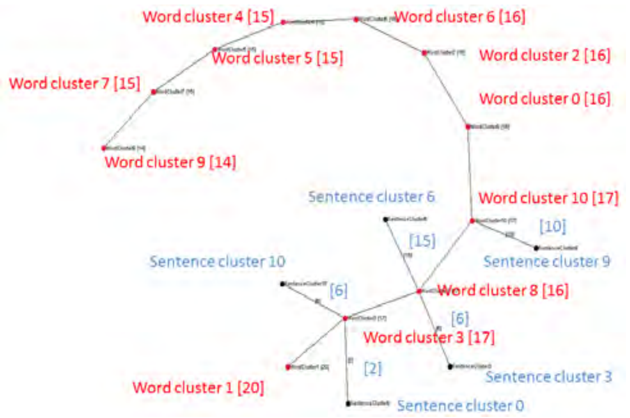

Fig. 7. The result of work of system for the text with weak semantic relations – semantic network of the document. WordCluster corresponds to clusters – stemmata, SentenceCluster corresponds to clusters – sentences.

TABLE II. PROVISION OF TEXT COHESIVENESS.

n	***n***	***n***	***n***	***n***	***n***	***n***	***n***	***n***	***n***
0	0	0	0	0	0	0	0	0	0
z	***z***	***z***	***z***	***z***	***z***	***z***	***z***	***z***	***s***
1	1	1	1	1	1	1	1	1	0

VI. CONCLUSION

There was formed and realized, in the course of the work, at the software level the approach to the applied creation of the semantic nets of the natural language text. The main advantage of the described model is a high degree of adaptability, expressed in the absence of the need for preliminary linguistic annotation of the working set of texts or adding in the system basis some set of linguistic rules. The made researches show that in spite of the preliminary absence of any deep semantic knowledge, the system is resistant to the frequency actions from the side of automatically generated text with low semantic indexes and it is oriented precisely to the processing of semantic relations in the text. The created approach provided the basis for the system of automatic determination of the text semantic connectivity, the work results of which confirm the applied applicability of the developed method for the semantic net construction to the tasks of program processing of the complex semantic text structures regardless of the natural language.

REFERENCES

[1] A. N. Shevtsov, S. I. Sorokin, and Y. O. Mamadkulov, “System of synthesis of educational tests on the basis of formal grammars,” journal "Software programs and systems", no. 2(102), p.181-185, 2013.

[2] N. I. Gurin, and Y. A. Zhuk, “Semantic network of an electronic workbook for dialog with virtual teacher,” international scientific and technical Internet conference “Information Processing Technologies in Education, Science and Production”, Belorussian State Technological University, Minsk, 2015.

[3] M. Yu. Gubin, V. V. Razin, and A. F. Tuzovsky, “Application of semantic networks and frequency characteristics of texts on natural languages for the creation of semantic metapopsis,” Problems of Informatics, p.59-64, 2011.

[4] I. I. Yusupova, and M. M. Gayanova, “Semantic Networks and Producing Models for the Analysis of University Educational Programs,” Newsletter of Ufa state aviation technical university, pp. 120-126, 2006.

[5] N. G. Zagoruiko, A. M. Naletov, and I. M. Grebenkin, “On the way to automatic construction ontology,” Materials of international conference "Dialog", 2013.

[6] Pismak A.E., Kharitonova A.E. (2016) // The method of automatic formation of a semantic network from weakly structured sources // . Nauch.-tekhnich. vestn. ITMO [Scientific and Technical Journal of Information Technologies, Mechanics and Optics]. 2016, vol. 16, no. 2, p. 324-330.

[7] Tarasov D. S. (2015) - Natural language generation, paraphrasing and summarization of user reviews with recurrent neural networks [Text] /Tarasov D. S./ "Computer linguistics and Intellectual Technologies", No14(vol. 1), 2015, p.607-614//Materials of international conference "Dialog", 2015.

[8] O.S. Volkovsky, Y.R. Kovylin. Computer system of automatic determination of the text coherence [Text] // System Technologies; Regional Interuniversity Collection of Scientific Papers. -Release 1 (112) 2017. - Dnipro, 2017.

[9] O.S. Volkovsky, Y.R. Kovylin. Analysis of the modern approaches to the task of automatic text generation in the natural language [Text] // System Technologies; Regional Interuniversity Collection of Scientific Papers. - Release 1 (100) 2016. - Dnepropetrovsk, 2016.

[10] N.N. Leontyeva. (2006) Automated comprehension of texts: Systems, models, resources. [Text]/N.N. Leontyeva//Moscow – 2006.