UDK 004.738.52

O.S. Volkovskyi, Y.R. Kovylin

COMPUTER SYSTEM OF AUTOMATIC DETERMINATION OF THE TEXT COHERENCE

Abstract — Stipulated and implemented adaptive model of formation of the semantic text net. An approach to the rating of numerical characteristics of the semantic properties of the text is considered. There has been developed an algorithm for the normalization of obtained data for its usage during the studying of neural net in order to differentiate the coherent text from semantically «noisy» one. On the basis of described models and algorithms of text processing, the software application was implemented and tested.

Introduction. Field of automated text processing contains a lot of tasks – starting from the simplest tasks, such as similarity matching or automated abstracting and finishing with the extremely difficult, such as text production and translation process. Under the complexity, we mean profundity of semantic analysis that must be done for the proper system operation. It provokes certain algorithmical difficulties, as a native natural language is non-systemic and it lends itself for machine realization with difficulties. The popular way to solve this problem is a description of rules and exceptions systems that are biased to the certain knowledge database – the typical example is some translation systems that store both – dictionaries and rules of sentences formation during the translation process. Thus, such approach does not solve the issue concerning the decision – making about the text coherence, as nowadays there is no adaptive mechanism of automated description of the document semantic structure – the existing decisions are based on the pattern, designed before structures of documents, and consequently, systems of this kind are dealing with documents of one thematic scope.

Challenge problem. Under the text coherence, we will understand the level of semantic correspondence of the text elements towards each other; so, the topic presentation must be performed proportionally throughout the whole document, taking into consideration the fact that a document must be devoted to one topic. Within this framework, it is

necessary to perform the adaptive algorithm of receipt of document's semantic properties in the numerical form that will allow to estimate text automatically in the context of its coherence and that would not be biased to the certain subject and big knowledge database.

Analysis of the latest researches and publications. The existing models of text coherence based on the native language are produced from the statistical data analysis, for example, in the work of [1] that leads to the decrease of the model adaptiveness because of the fact that the obtained data was used only for the same subject area, from where it was taken. Another way to solve this problem is the usage of vocabulary reference pattern and corresponding rules of their usage. [2]. The complexity of such systems is a processing of pattern knowledge databases, because, in order to obtain such results, a developer must not only create and place linguistically a huge corpus of data, but also, he needs to develop the navigation mechanisms throughout the whole corpus and the adaptive enclosure of new data boxes. The greatest project that would solve the most part of issues is a semantic web [3], that mean that each web – page that is located in the Internet will contain the certain semantical annotation. Thus, even if we ignore the fact of complexity during the process of such annotation introduction on the part of user and the absence of mechanism that can look for mistakes in the created annotation, nowadays there are no conditions to implement such approach in the nearest future worldwide.

Objective of the research. To create the approach for the formation of the semantic net of the document. To develop methods in order to obtain numerical analogues of the semantic characteristics of the documents. To test and endorse the obtained results in the form of application systems for evaluation of text coherence.

Statement of basic materials. The first stage that must be passed by any developer of systems for automated processing of texts is a syntactic analysis. At this stage, there is a detachment of sentences and words of the analyzed text. In addition to it, there is a contraction of many words due to stemming and withdrawal of the auxiliary parts of speech. For this purpose, each pair of words is being cut of endings pursuant to the Porter's algorithm, and then the distance of Levenshtein is being subtracted for the obtained results. If the meaning is more or equal to the length of the most general part of the analyzed words, it is

considered to be that stem has been found and each word is being changed by the revealed general part.

Next step of the syntactic analysis is a definition of the language parts stem in order to withdraw words without any information (such as auxiliary parts of speech) from the process of semantic analysis. For this purpose, the system has a marked sample in size of fifteen thousand of words and correspondent parts of language that serves as a studying corpus for the Naive Bayes Classifier, where the classes are parts of speech and the corresponding meanings to the class are two or three last letters of the initial word and the ending obtained pursuant to the Porter's algorithm. Each word from the analyzed text is being classified on the model and if the forecast states that this word is not informative, it will be deleted.

A concluding stage of the syntactic analysis is a measurement of stems, so that each stem has a number of repetitions in the text and measurement of the sentences, where the weight function of the sentence means total weight of all stems in the sentence.

A test analyzed in this way, must pass the stage of frequency response analysis, so that the text data will have the equivalents in the numerical characteristics. In order to achieve such result, it is offered to compose the matrix, which lines correspond to the sentences, the columns correspond to the stems and the meanings are numbers of stems in the sentence. After we obtain such matrix, we need to perform on it a process of singular value decomposition. Singular value decomposition is steady, it is possible to take away those meanings of left and right matrix that corresponds to the low singular meanings and to leave only two biggest meanings, after that, it is possible to use them as the coordinates for reflection on the two-dimensional surface. The obtained results are reflected in the figure 1 and figure 2.

The next step is to cluster points for stems and sentences under the algorithm k-means. The number of clusters for stems and sentences *cl* is being indicated pursuant to the formula (1):

$$cl(W, W_U) = \frac{count(W)}{count(W_U)}, \quad (1)$$

where W means words, W_U means stems. Centroids of cluster – stems are positions of stems with the most frequency in the text, that is being revealed pursuant to the formula (2):

$$Cst(W_U) = \max(W_0 \ldots W_{cl}), \quad (2)$$

where $W_0 \ldots W_{cl}$ are weights of stems. Centroids of cluster – sentences are positions of sentences with the biggest total weight of stems that is being revealed pursuant to the formula (3):

$$Cs(W_S) = \max\left(\sum_{i=0}^{SN} W_i\right), \quad (3)$$

where W_S is a sentence, W_i is a stem weight in the sentence, SN – is a stem numbers in the sentence.

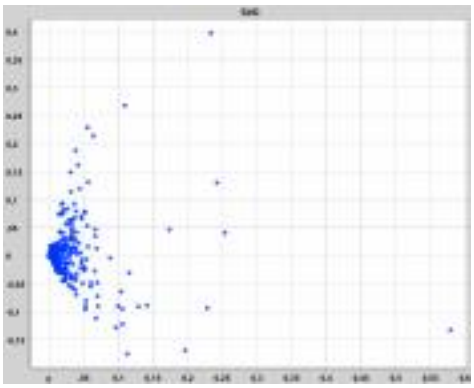

Figure 1 – Stem projection

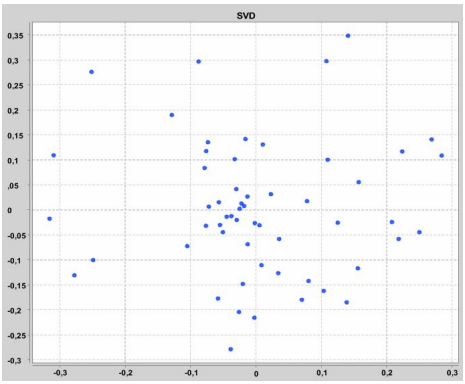

Figure 2 – Sentences projection

On the basis of the points positions of each cluster – stem in accordance with the Jarvis's Algorithm, the outline of convex figure is being created. The obtained results reflected in the figure 3 for stem and figure 4 for sentences.

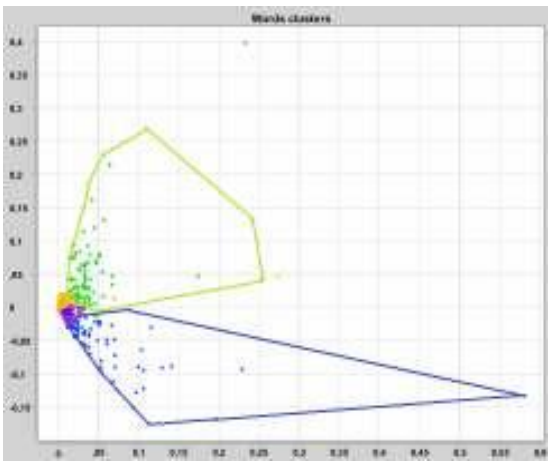

Figure 3 - Convex figures of the clusters – stems

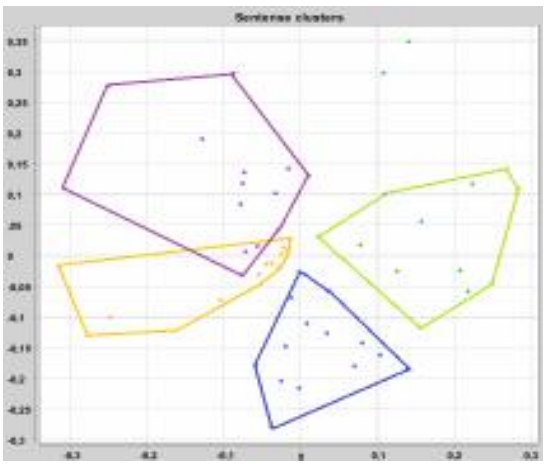

Figure 4 - Convex figures of the clusters – sentences

For each cluster – stem, the weight must be stipulated – number of stems in it, on this basis, there has been built a semantic graph of

clusters connection in the descending order of their weight. For each figure of clusters – stem obtained pursuant to the Jarvis's Algorithm, there must be checked the hit of points that form each cluster – sentence. If it is possible to find such points - a cluster of the sentence joins with the cluster – stem in the net, where the link weight is a number of points that exist in the outline of the cluster – stem. The result of system operation is reflected in the figure 5.

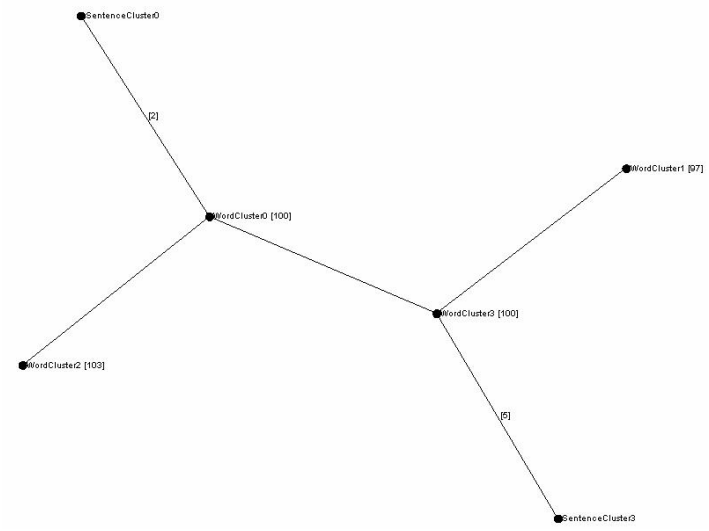

Figure 5 - Semantic net of the document

The obtained semantic net may be used to take the numerical data that characterize semantic properties of the document and may be used for the automatic determination of the text coherence. During the researches, it was stipulated that such characteristics include total amount of stems, quantity of all words, quantity of cluster – systems that have a connection with cluster – sentences and total amount of cluster – stems.

The obtained data is being transferred to the entry of asynchronous neural network that is on the basis of data of sample studying corpus takes a decision concerning the text coherence. It is necessary to take into consideration that the described semantic characteristics depend on the text size, so the taken data requires previous normalization. For this purpose, there was composed a corpus of eighty texts concerning informational technology, astronomy, and incoherent texts that were received due to the services of frequency-response auto generation, each of them is being characterized with two meanings – normalized text size W_N, obtained pursuant to the formula (4):

$$W_N = \frac{W_i - W_{\min}}{W_{\max} - W_{\min}}, \quad (4)$$

where W_i means the total amount of the words, W_{min} and W_{max} mean the smallest and the biggest amount of words in the studying corpus and normalized semantic meaning S_N, obtained pursuant to the formula (5):

$$S_N = \frac{W_U}{W} \bullet \frac{CW_C}{CW}, \quad (5)$$

where W_U means the total amount of the stems, W – is a total amount of words, CW_C - is a quantity of clusters – stems that have a correlation with clusters – sentences, CW – is a total amount of clusters – stems. The obtained result is a training sample for the neural net.

In order to test the system, there has been composed a sample of 20 texts, as incoherent (auto generated) and real scientific tests on the following topics: astronomy, informational technology and economics. Besides, a sample includes a text, which was formed from the different coherent parts of the text of the same topic, but in general, it is semantically incoherent. The results of the text processing are reflected in the schedule 1, where «n» corresponds to the auto generated text, «z» corresponds to the coherent text; «s» corresponds to the text composed of different parts, 1 means that a forecast indicates text coherence, 0 means that a forecast indicates the texts incoherence.

Table 1

Forecast of coherence

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

Conclusions and directions for future research. Pursuant to the results of the work, there has been developed an approach to create a semantic net of the text that may be used for taking quantity specifications of the semantic properties of the document. Contrary to the existing approaches to the automated creation of the semantic picture of document, a described methodology helps to obtain the semantic structure of the text without any additional linguistic knowledge that means its previous linguistic annotation, sample subject knowledge databases or special systems of linguistic rules. There has been conducted a research of the influence of quantity specifications of the semantic net on the meaning of text coherence pursuant to which results, there were de-

tected not only considerable semantic characteristics, but developed mechanisms for their normalization. On the basis of the obtained data, there has been created an application system of the automated detection of the text coherence, which testing results indicate a success of the developed model usage in the process of solving the target task.

REFERENCE

1. A.N. Shvetsov, S.I. Sorokin, Yu.O. Mamadkulov – Educational Test Synthesis System based on the formal grammar // RI "Centerprogrammsystem" - magazine "Software Products and Systems", No.2 (102), 2013, pp.181-185.

2. N.I. Gurin, Y.A. Zhuk - Semantic Network of the electronic textbook for a dialogue with a virtual teacher // Materials of the international scientific and technical online conference "Information technologies in education, science and production" // Belarusian State Technological University, Minsk, 2015.

3. Haarselv V., Moller R. – A core inference engine for the Semantic Web// Proc. of the 2nd International workshop on evaluation of ontology-based tools (EON-2003) – Florida, USA, 2003.

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