МАТЕМАТИЧНЕ МОДЕЛЮВАННЯ В ПРИРОДНИЧИХ НАУКАХ ТА ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ

DOI:
UDC 004.738.5

Y.R. Kovylin, Post-graduate student of the Department of Computer Science and Information Technologies, Oles Honchar Dnipro National University, kovilin.yegor@gmail.com
O.S. Volkovskiy, Cand. Sci. (Tech.), Assistant professor of the Department of Computer Science and Information Technologies, Oles Honchar Dnipro National University, Dnipro

COMPUTER METHODS FOR COMPILING AN ENTRY OF EXPLANATORY COMBINATORIAL DICTIONARY BELONGING TO "MEANING↔TEXT" THEORY WITHIN THE TASK OF TEXT AUTOMATIC GENERATION

The objective of this article is developing a new computer model of explanatory combinatorial dictionary for intellectual systems of text generation based on Meaning-text theory (MTT) and solving problems of model software implementation by means of mathematical and algorithmic description of article paragraphs and preliminary formation and preparation of knowledge bases aimed at improving system adaptivity. Researches of theoretical insights into the issue of NL description as well as researches of existing software text generation facilities were performed. Assessment of considered methods was performed and choice of these methods was grounded taking into account their applicability to solving the matter of text generation. On the basis of the research results the most appropriate theory of language formal description was chosen. Main disadvantages of the selected theory were defined for developing applicable software systems of scientific text generation. With a help of statistical analysis methods and instruments of artificial intelligence and algorithm of solving theory scientific problems in the context of forming question-and-answer system of automatic text synthesis.

***Keywords**: Computer linguistics, quasi-abstracting, «Meaning-text» theory.*

Метою статті є розробка комп'ютерної моделі толково-комбінаторного словника для інтелектуальних систем генерації тексту що базуються на теорії «Смисл-тескт» та вирішує проблеми програмної реалізаціїї за допомогою математичного і алгоритмічного опису пунктів статей без попереднього формування та підготовки бази знань для підвищення адаптивності системи. Проведено дослідження, як теоретичних розробок опису природної мови, так і існуючих прикладних програмних реалізацій систем генерації текстів. Виконано оцінку і обґрунтований вибір розглянутих методів з точки зору їх застосовності до задачі генерації текстів. На основі аналізу результатів дослідження була обрана найбільш підходяща теорія формального опису мови. Визначено головні мінуси обраної теорії для завдання розробки прикладних програмних систем генерації наукового тексту. За допомогою методів статистичного аналізу та інструментів штучного інтелекту розроблено алгоритм вирішення наукових проблем теорії в рамках завдання побудови запитно-відповідальних систем автоматичного синтезу тексту.

***Ключові слова**: Комп'ютерна лінгвістика, квазіреферування, теорія «Смисл-текст».*

Problem statement

The issue of automate text processing (ATP) refers to a great number of scientific matters which are first of all related with problems of algorithmic implementation of such products. Development of ATP applied software systems presupposes choice of a particular mechanism of natural language (NL) description and implementation (methods which are accessible for computer). But language is quite an unformalized system which is characterized through irregularity and non-uniformity of its rules. And the main problem consists in complexity of describing semantic characteristics of text on the level of algorithmic representation. NL is not only a set of words based on some grammar components (obtaining a really intelligent text is a top-priority task of ATP). And this fact brings many developers to a necessity of taking into account semantic relations not only between separate words but also between separate sentences and even between separate documents. Usually text semantics and text computer-based perception of text mean the following: if we enter a certain text to the computer memory and print it with a help of a printer semantics is not meant; but if this text is processed and in the result of this processing the user receives a new text being clear and adequate for him/her (for example translation into another language) we can talk about computer-based semantic perception of text. In this context intellectual generation of texts is the most complicated because during semantic processing the template of text synthesis is vaguer in comparison with, for example, automatic translation. The gap between language linguistic description and its applied implementation makes this process even more difficult — linguistics is first of all focused on description of language nature by means of notions taken from insufficiently formalized sciences (such sciences as psychology, philosophy, anthropology etc.). During realization of automated text processing systems developers when using computer science instruments have to adapt these instruments to working with NL and to solve problems which do not have anything to do with the classic linguistics. And that gave rise to such hybrid sphere of science as computer linguistics which is already aimed at mathematic modeling of NL. And that's why an important task consists in performing a grounded choice of NL model which is not only of theoretical importance but which also gives an opportunity to create an applied software implementation for up-to-date computers

An important characteristic of NL is inflection — a property of language variability which depends on the range of multiple lexical endings for various parts of speech. This property causes a direct influence on descriptive complexity of NL computer model — the greater is the number of inflexions in a language, the freer is the word order in sentences of this language and the less formal is this language. From this point of view the English language is much easier for creating a computer model as far as it has a strict word order in sentences and poor inflexion of endings. Indo-European languages (Slavic languages belong to this group of languages) are characterized through a free order of semes and a complex system of inflexions and that is why the western and the domestic computer linguistics were developed with a use of different routes of development. In western countries Chomskiy's theories of grammar components became popular [2], and the domestic computer linguistics is based on Melchuk's semantic theory MTT presupposing that meaning of a text is more important than its grammar. Such scientists as Leontyeva, Apresyan, Bolshakov [3—5] have performed multiple attempts to modify and to develop applied variants of MTT model but as far as the theory was initially created for the process of test automatic translation implemented systems were also intended for translation from Russian into English (French). But neither of these systems solved the problem of explanatory combinatorial dictionary. Text synthesis model was viewed from many points and that`s why an interesting task is to apply this model to solving the task of intellectual generation of text results obtained with a help of question-and-answer system and that requires a certain modification of the initial theory. In this work we are going to look through main approaches to description of NL in systems of generation of comprehended texts including inflectionally rich text information with some semantic data; and we also propose a model for removing restrictions from applied algorithmic implementation of the chosen theory.

Analysis of recent research and publications

The up-to-date computer linguistics includes several classes of text generation systems which differ in complexity of data processing and in complexity of their intellectual component. Provisionally these systems may be divided taking into account language models these systems are based on: Chomsky's generative grammars, semantic network and instruments of neural networks. As far as the necessity of applied implementation in this case is more important than the theoretical component let's look through concrete software systems in each of the mentioned classes. Let's start with the program of generating test tasks for distant learning of students; this software product is based on paradigms of Chomsky's formal grammars [2]. Generative grammar of components is based on the axiom about the phenomenon of language competence which is presented as human ability to comprehend and to understand natural human speech irrespective of the language. On the ground of this generative grammar sets a task to model this ability by means of forming correct sentences and by means of using a definite final set of rules, alphabet and sentence symbol which may be used for expanding schemes of sentence structure (immediate constituents) with a help of special grammar rules. Theoretically the great number of immediate constituents is not restricted in any way and is unlimited. In practice the language itself, the subject area, the body of the text and computer capabilities significantly restrict the number of immediate constituents. Semantic network technology became really widespread in the sphere of ATP. This technology is the next step in development of text processing. Semantic network is a graph with semantic units in peaks and the arcs describe semantic relations between them. Semantic units are usually perceived as separate words, sentences and even separate documents. Practical application of semantic network for solving the task of text generation is well illustrated in work [6] — automatic consultation system. Developers set themselves a task to generate knowledge bases for separate subject areas in order to ensure dialog with users taking into account respective issues of these subject areas. Semantic network is proposed to be used for storage of extracted knowledge on the basis of a definite text body which is presented as sets of hackneyed phrases prepared in advance (answers).

Up to date the most advanced instrument applied for solving the task of text automatic generation is presented as applied methods of artificial intelligence (implementation of ATP with a help of neural networks). Up to date artificial neural networks are broadly used for solving various applied tasks of artificial intelligence (AI) including tasks of language automatic processing. As to the issue of generating new texts developers and scientists usually use neural networks of three types: the simple recurrent network for processing such series of units as sentences; the recursive auto-associative memory for processing linguistic structures presented in form of trees and Kohonen Self-Organizing Maps used for clustering such representations. In order to assess quality of methods of using neural networks for solving the task of text generation let's look through the work [8] where a recurrent network is used for describing products of a certain Internet shop. As we can see the results are quite ambiguous. The main advantage of this approach consists in complete automation of text generation process, high level of system adaptivity and low costs of its adjustment and introduction. But there are some evident problems of semantic garbage such as "beautiful speed, responsive screen, working day". The reason of that consists in the fact that despite apparent intellectual processing the system does not understand the sense of what it describes and so it generates units taking into account exclusively preliminarily prepared templates (teachers).

After viewing main approaches to the task of text generation and main applied implementations of these approaches we can see their weak applicability to solving the task of intellectual text generation. As to Chomsky's generative grammars we should point out that despite an opportunity of quick generation of texts (system tests) and high flexibility of introduction the system of distant learning in particular and Chomsky's language model in general do not solve the problem of understanding the NL and intellectual text generation. In order to prove this we should turn to the theory and practical application of generative grammar. Many linguists were against application of the generative model as the main model for Indo-European languages. So, M.M. Mozgovoy states the following: "…Chomsky's grammars are first of all intended for describing structure of sentences. And the matter of describing meanings of separate words (this matter is not less important) remains beyond opportunities presented by these grammars".

As for semantic networks their absolute advantage consists in simplicity of their implementation. But at the same time this approach has a number of disadvantages. The most significant disadvantages include: weak system adaptivity and difficult process of network structural changes in case of subject area changes — in such cases preliminarily prepared templates may work against the system. In addition to that decision making in conditions when the received information does not correspond to the template is not controlled or processed in any way. As far as approaches based on semantic and lexical analysis are quite applicable for relatively simple linguistic tasks (just like rubrications with a-priori known rubrics) such control and processing is of no use. But if automatic generation of texts is meant there is an evident necessity to have a certain intellectual system component responsible for taking decisions and performing analysis of text semantic constituent.

It is just the sphere where neural networks are used. Up to date attempts to combine the template algorithm and the component presented in form of neural networks give quite good results. But even despite this fact computer does not in fact perceive text semantics — and the above presented results illustrate this fact perfectly well. In addition to that systems of automatic translation may be taken as an example. In this case application of neural networks did not lead to receiving applicable results. Often in attempts to improve operability of such systems developers wend the way of complicating the network depending on the task or the subject area. And that invariably lead to losses in system adaptivity and transferability. For solving this problem within this work we propose to adapt the method of NL description developed by Melchuk and named MTT for automatic generation of texts. MTT separates semantics from syntax and at the same time this theory insists on its scientific description. The main advantage of MTT consists in initial orientation of this theory not on grammar components but on meaning being the main object of the model. And that is well suited for solving the task of text synthesis. "…As far as we know text synthesis performed in accordance with an arbitrarily assigned semantic network was seriously analyzed exactly within MTT model…"[4]. Text synthesis requires description of underlying semantic relations (super-phrasal relations) which are going to model the process of text intellectual understanding. As it has already been mentioned such semantic relations are not taken into account either in Chomsky's grammars or in semantic and neural networks as far as rules of the generative model are aimed at solving grammatical ambiguities and network models work in accordance with predefined templates. The main disadvantage of MTT is in explanatory combinatorial dictionary. The mechanism of this dictionary consists in manual description of each word in the subject area and its semantic relations. It presupposes that each seme is going to receive the complete description of its properties — starting from lexical characteristics and up to such high semantic levels as related idioms. Compilation of such a dictionary is quite a labor-intensive task. This may be explained by the fact that up to date there are no behavior scenario algorithms for systems based on MTT in case of analysis of an unknown word. And this is connected with the fact that structure of an article is too complicated and if there is an opportunity for automatic choice of morphological markers with a certain admissible error other points (such as for example idioms) cannot be used automatically without special knowledge So, if MTT is viewed as a language model of text generation system then top priority scientific problems are presented in form of issues related with learning the system, overcoming limitations of knowledge bases and flexible adjustment of system operation when new subject areas are included. In this work we are going to redefine the model of dictionary in a way giving an opportunity for its automatic filling on the basis of text compression algorithm with a help of weight coefficients and methods of seme classification depending on lexical and morphological characteristics of these semes..

Formulation of the research purpose

In the process of our work we are facing several tasks. To verify the work of the MTT, the problem of semantic quasi-reframing was chosen. It is necessary to develop an applied methodology for quasi-abstracting both from the point of view of frequency analysis and a certain program model of the text that allows obtaining results similar to the MTT for quasi-abstracting, but with no applied constraints on the theory. The both processes should comply with the applied program implementation and they should be completely automated.

Presentation of the main material

When viewing the system of text generation on the basis of MTT the top priority issue consists in the structure of the knowledge base (KB) used for selection of text data for formulation of the respective system response. The initial Melchuk's theory defines this KB as explanatory combinatorial dictionary. In our research KB will be presented in form of a semantic model of the document. In comparison with the widespread frequency methods of reflowing, the use of the dictionary and the theory would allow us to weigh the sentences not only having a frequency connection with the most frequent words in the text, but also taking into account their semantic connection with other concepts. Thus, in the answer there will be sentences not containing in themselves words with the greatest frequency, but having thus deep semantic communication with a theme of the document. On this stage it is important to understand that usage of a preliminarily specified morphological dictionary will significantly reduce adaptability and tolerability of the system. Instead of that we propose create the approach for the formation of the semantic model of the document.

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 twenty 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.

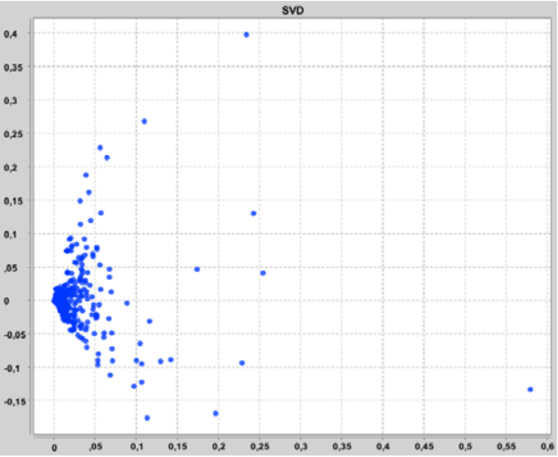

Fig. 1. Stem projection.

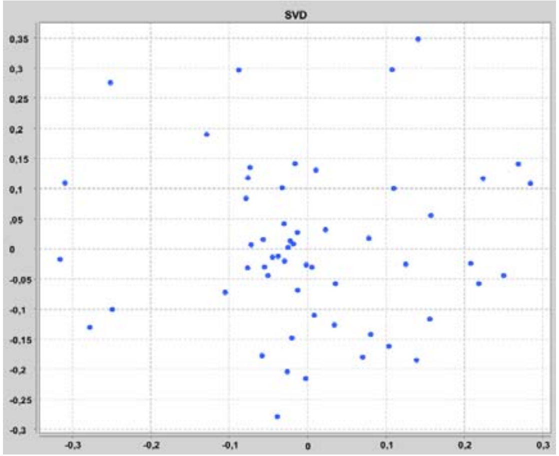

Fig. 2. Sentences projection

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):

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

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):

$$Cst(W_U) = \max(W_0 .. W_{cl}), \tag{2}$$

where $W_0 ... W_{cl}$ 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):

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

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

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.

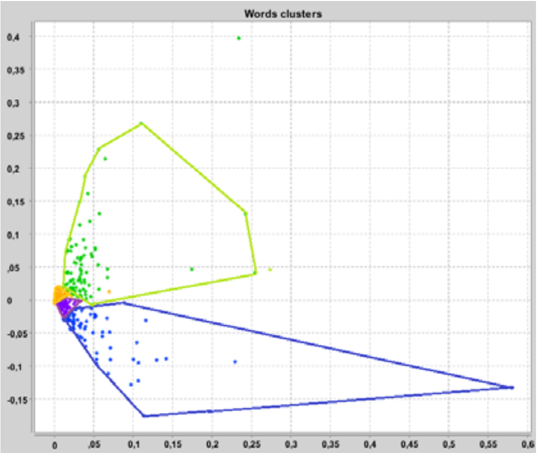

Fig. 3. Convex figures of the clusters-stems

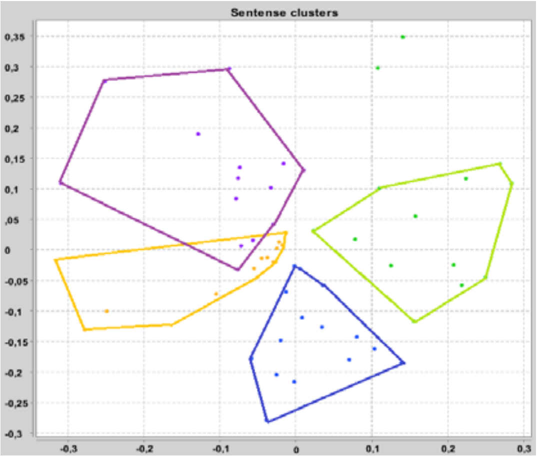

Fig. 4. Convex figures of the clusters-sentences

The obtained semantic net may be used to take the numerical data that characterize semantic and morphological properties of the document and may be used, for example, for the automatic semantic quasi-abstracting. When a document arrives at the system, each word cluster checks the occurrence of the stemma with the maximum weight in the document. If this occurrence is found for the current cluster of stems, then the clusters associated with this cluster become a candidate for inclusion in the resulting response. If there are several such connected clusters, then a cluster with the maximum connection weight enters the set of candidates. In addition to the text of the proposal itself, the cluster contains data on the offer number and its weight relative to the document being processed. Such an operation is carried out over document, as a result of which we receive a multitude of sentences of candidates for inclusion in the response. For their normalization, the weight of each candidate is divided by the amount of words from its source text. The resultant answer includes the candidates with the maximum normalized weight, sorted by their original number in the text (and if the numbers are the same, then by weight). The size of the resulting response is defined as the ratio of the total number of words in the body to the total number of sentences.

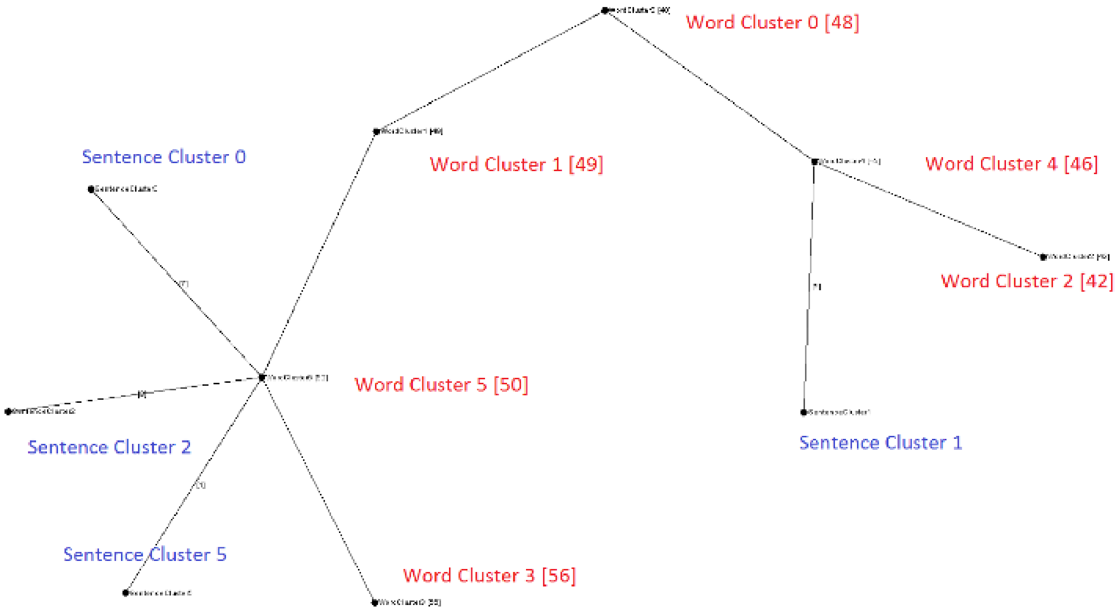

Fig. 5. Semantic morphological net of the test document

To verify the semantic properties of the described model, we also additionally compiled an algorithm for frequency quasi-abstracting. Input text document are subjected to lemmatization (a kind of morphological analysis presupposing leading stems to their initial dictionary forms). In the result of lemmatization separate word forms are deprived of their endings and brought to the standard form which gives an opportunity to obtain more accurate results during semantic quasi-abstracting. Each sentence is subjected to the procedure of removing stop-words. All stop-words are taken to a special dictionary. The next stage consists in stemming operation with all remaining words in accordance with Porter's algorithm. Stemming is a process of finding stems of words (which may not always coincide with the morphological stem). The operation of preliminary lemmatization gives an opportunity to reduce the number of mistakes (availability of mistakes can be explained by imperfection of Porter's algorithm). The algorithm includes four steps. At the first stage the maximal formative suffix is removed; at the second stage letter "n" is removed; the third stage words are deprived of their word formative suffixes; and at the fourth stage words are deprived of superlative suffixes.

When viewing algorithms of ending exclusion it should be noted that words are often reduced to a greater extent than it may be needed. And in addition to that these algorithms are not resistant to loss and substitution of letters in roots and suffixes during word formation. Lemmatization reduces the number of such mistakes but it does not give an opportunity to achieve accuracy required for semantic analysis. And that's why each pair of words already subjected to lemmatization and stemming are then subjected to the procedure of finding the maximal largest part of word. Complexity of this stage consists in assessment of suitability of the determined common part. In order to understand the level of suitability of the determined common part we use mechanism based on assessment of word lengths. According to the following formula:

$$EQ = Q * \max(wl_i, wl_j) \,, \tag{4}$$

where Q — semantic coefficient of equivalence, $max(wl_{i,}\ wl_j)$ — function of finding the maximal length out of lengths of two words ($wl_{i,}\ wl_j$) we obtain value EQ. If the value of EQ is less than the length of the determined common part then it is suitable for semantic analysis and each word of the investigated pair of words is substituted with the determined common part. Coefficient Q shows the percentage of necessary coincidence of the determined part with the maximal length of the word from the investigated pair of words. It was determined experimentally that the optimal value of Q is 0.4. So, that is the way we use for finding bodies of words in concrete texts which can be used for weighing. Weight of each sentence is calculated according to the following formula:

$$W_S = \sum_i^N F * Q_i + H * \frac{WC}{SC}. \quad (5)$$

Where: $F*Q_i$ — frequency of word in the text, N — number of words in the sentence, H — number of hint-words in the sentence, WC — number of words in the text, SC — number of sentences in the text. In the result of that we obtain a set of pairs-sentence and its weight. The last task is to compile the final report. For this purpose we calculate the level of the initial text compression and so the number of sentences in the final report becomes clear: out of the great number of sentences we choose sentences of the highest weight and we include these sentences to the final report in the order they appear in the text. Sentences tagged as links with the previous sentence not included to the report are also removed. This procedure is used for removing semantic gaps in the report.

The result of the obtained semantic model of the document is shown in fig.5. The resulting graph contains six vertices of the stems of which only two have links to clusters-sentences. Table 1 contains the result of processing incoming text using the semantic model of the document and the result of processing the same text by the frequency analysis method.

Table 1. Results of quasi-abstract

Result semantic quasi-abstract	Result frequency quasi-abstract
A specific apple is generally an apple, an apple in general, and any apple in general is a fruit. That is why examples of classes in the textbooks on object-oriented programming so often mention apples and pears. In the future, the words "class", "object", "interface" and "structure" will be used in their special values specified in the framework of OOP. object-oriented program with the use of classes, each object is an "instance" of a particular class, and other objects are not provided. Static fields exist in one instance for the entire program (or, in a more complex version, in one instance per process or thread / thread). Regular fields are created one copy for each specific object - an instance of the class. The algorithms themselves, that is, the actual program code that will perform all these calculations, are not specified by the interface, it is programmed separately and is called the implementation of the interface. Program interfaces, as well as classes, can be extended by inheritance, which is one of the important means of reusing ready-made code in OOP. The inherited class or interface will contain everything that is specified for all its parent classes (depending on the programming language and platform, they can be from zero to infinity). At the same time, inheriting the class, we automatically inherit the ready-made code for the interface (this is not always the case, the parent class can require the implementation of some algorithms in the child class without fail). The words "private" and "public" in this case are so-called "access modifiers". There is a basic rule: nothing in one class can see private elements of another class. Inheritance by private type makes all members of the parent class (including public and protected) private members of the successor class (C ++); protected (protected, internal member of the class hierarchy) - member access is allowed from methods of the class in which this member is defined, as well as from any methods of its heir classes.	A class is a software element that describes an abstract data type and its partial or complete implementation. Other abstract data types - metaclasses, interfaces, structures, enumerations, - are characterized by some of their own, other features. Along with the concept of "object", the class is a key concept in OOP (although there are classless object-oriented languages, for example, Self, Lua; see Prototype Programming for details). The essence of the difference between classes from other abstract data types is that when specifying a data type, the class defines both the interface and the implementation for all of its instances simultaneously, and the call of the constructor method is mandatory. In practice, object-oriented programming reduces to creating a certain number of classes, including the interface and implementation, and their subsequent use. Human classifications in zoology, botany, chemistry, machine parts, carry the basic idea that anything can always be represented as a particular case of some more general concept. Everywhere further, the words "class", "object", "interface" and "structure" will be used in their special values specified in the framework of OOP. In an object-oriented program using classes, each object is an "instance" of a particular class, and no other objects are provided. That is, the "instance of a class" in this case does not mean "an example of a certain class" or "an individual class", but "an object whose type is a class". In this case, in the different programming languages, the existence of some other data types whose instances are not objects is allowed (that is, the language determines whether objects are things such as numbers, arrays and pointers, or are not, and, accordingly, whether there are such classes as "number", "array" or "pointer", whose instances would be any particular number, array or pointer). When using classes, all elements of the program code, such as variables, constants, methods, procedures and functions, can belong (and in many languages must belong) to a particular class.

Conclusions and prospects for further research

On the ground of the performed researches we can conclude that application of Melchuk's theory MTT for solving the task of automatic generation of text responses of intellectual inquiry system is more than justified. Thanks to using this theory as a language model the described approach will give an opportunity to synthesize considered responses (answers) to definite questions (documents) of the user. The word with the maximum weight in the received document was the word "class". Proceeding from this, we can conclude that the model created is supported by semantic rather than frequency links in the text. The text given in table 1 for semantic quasi-abstract contains sentences whose connection with the original word is understandable only from the point of view of semantics. We are talking about such sentences like «a specific apple is generally an apple, an apple in general, and any apple in general is a fruit» — a clear example of the description of the concept of class, or like «the words "private" and "public" in this case are so-called "access modifiers"» — the concept of encapsulation, inextricably linked with object-oriented programming.On the other hand, the text in table 1 for frequency quasi-abstract contains only sentences containing the original search word, supporting only the frequency links.The described semantic model of the document is similar to the model of the semantic network, but it has some number of scientific differences. For their description, let's compare the existing developments in this field with our campaign. 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 [7]. For example, the word "burn" can be described as (fire, action). The practical application of this technology is described detailed in the work of [7], on the basis of which it is created a semantic metadescription 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 further development of the semantic nets technology received in work of [8]. 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 [9]. 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.

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 [7] a system, which was initially based on the articles of the language 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 [8] 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 [9]. 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.

List of reference links

[1] Mel'chuk A. Exact Methods in Linguistic Research / Mel'chuk A. // Journal of Linguistics – Los Angeles: University of California Press, 1963. – 184.p.

[2] Chomsky N. Human Language and Other Semiotic Systems / Chomsky N. // Semiotica, Volume 25, Issue 1–2, p31–44.

[3] N.N. Leontyeva. Semantic Dictionary for Text Understanding and Summarization /N.N. Leontyeva // International Journal of Translation. New Dehli. 2003. – 107–114. p.

[4] I.A. Bolshakov. The Meaning ↔ Text Model: Thirty Years After./I.A. Bolshakov, A.F. Gelbukh//J. International Forum on Information and Documentation, N 1, 2000.

[5] Y.D. Apresyan. Linguistic support of the ETAP-2 system/Y.D. Apresyan, I. M. Bulavskiy, L.L. Iomdin //M.: Nauka, 1989. – p. 294

[6] Gunin N. I. (2015) Semantic network of an electronic workbook for dialog with virtual teacher / Gurin N. I., Zhuk Y. A./ Materials of the international scientific and technical Internet conference "Information Processing Technologies in Education, Science and Production"/ ACM International Conference Proceeding Series, 2015.

[7] F. Lehmann. Semantic networks / Computers & Mathematics with Applications, Issues 2–5 – Great Britain, p. 1–50.

[8] S. Grünewald QNet: an agglomerative method for the construction of phylogenetic networks from weighted quartets/ Published by Oxford University Press – 27 p.

[9] P. Wojtinnek. Building semantic networks from plain text and wikipedia with application to semantic relatedness and noun compound paraphrasing // International Journal of Semantic Computing, 2012, p. 59–64.

КОМП'ЮТЕРНА МЕТОДИКА ФОРМУВАННЯ СТАТТІ ТЛУМАЧНО-КОМБІНАТОРНОГО СЛОВНИКА ТЕОРІЇ «СМИСЛ↔ТЕКСТ» В РАМКАХ ЗАВДАННЯ АВТОМАТИЧНОЇ ГЕНЕРАЦІЇ ТЕКСТІВ

Ковилін Є.Р., Волковський О.С.

Реферат

Мета. Розробка комп'ютерної моделі толково-комбінаторного словника для запитно-відповідальних інтелектуальних систем генерації тексту. Зняття проблем програмної реалізації моделі за допомогою математичного і алгоритмічного опису пунктів статей без попереднього формування та підготовки бази знань для підвищення адаптивності системи.

Методика. Проведено дослідження як теоретичних розробок опису природної мови, так і існуючих прикладних програмних реалізацій систем генерації текстів. Виконано оцінку і обґрунтований вибір розглянутих методів з точки зору їх застосовності до задачі генерації текстів. На основі аналізу результатів дослідження обрана найбільш доцільна теорія формального опису мови. Визначено головні мінуси обраної теорії для завдання розробки прикладних програмних систем генерації наукового тексту. За допомогою методів статистичного аналізу та інструментів штучного інтелекту розроблено алгоритм вирішення наукових проблем теорії в рамках завдання побудови запитно-відповідальних систем автоматичного синтезу тексту.

Результати. Було вибрано метод формального опису природної мови для завдання генерації текстів. Розроблено метод семантичного квазіреферування і оцінки вагових коефіцієнтів слів в тексті, на його основі створена програмна система автоматичного генерування рефератів. Запропоновано і досліджено метод зняття програмних обмежень на прикладну розробку

систем генерації текстів за допомогою розробленого методу автоматичної побудови семантичної мережі тесту.

Наукова новизна. Описано застосування теорій Мельчука «Сенс ↔ Текст» (ТСТ) для завдання автоматичної генерації текстів. Запропоновано метод подолання складнощів прикладної реалізації толково-комбінаторного словника в ТСТ і алгоритм отримання семантичних мереж тексту без попередньої семантичної розмітки.

Практична значимість. Обґрунтовано застосування ТСТ для формального опису мови в рамках завдання комп'ютерної інтелектуальної генерації текстів. За допомогою розроблених алгоритмів і структури статті словника подолані труднощі при програмній реалізації систем, які ґрунтуються на ТСТ. Створено алгоритм автоматичного семантичного стиснення текстів на природній мові та розроблена прикладна система реферування. Підвищено адаптивність і переносимість бази знань для систем генерації наукових текстів на основі ТСТ.

Література

1. Mel'chuk A. Exact Methods in Linguistic Research / Mel'chuk A. // Journal of Linguistics – Los Angeles: University of California Press, 1963. – 184 p.
2. Chomsky N. Human Language and Other Semiotic Systems / Chomsky N. // Semiotica, Volume 25, Issue 1–2, p. 31–44.
3. N.N. Leontyeva. Semantic Dictionary for Text Understanding and Summarization /N.N. Leontyeva // International Journal of Translation. New Dehli. 2003. – 107–114. p.
4. I.A. Bolshakov. The Meaning ↔ Text Model: Thirty Years After./I.A. Bolshakov, A.F. Gelbukh // J. International Forum on Information and Documentation, N 1, Mexico City, 2000.
5. Апресян Ю.Д.Лингвистическое обеспечение системы ЭТАП-2/Апресян Ю.Д. Богуславский И.М., Иомдин Л.Л. // М., Наука, 1989, 295 с.
6. Гурин Н.И. Семантическая сеть электронного учебника для диалога с виртуальным преподавателем / Гурин Н. И., Жук Я. А.//Материалы международной научно-технической интернет конференции "Информационные технологии в образовании, науке и производстве" // Белорусский государственный технологический университет, Минск, 2015.
7. F. Lehmann. Semantic networks / Computers & Mathematics with Applications, Issues 2–5 – Great Britain, p. 1–50.
8. S. Grünewald QNet: an agglomerative method for the construction of phylogenetic networks from weighted quartets/ Published by Oxford University Press – 27 p.
9. P. Wojtinnek. Building semantic networks from plain text and wikipedia with application to semantic relatedness and noun compound paraphrasing // International Journal of Semantic Computing, 2012, p. 59–64.