UDC 004.519.7 (045)
A. I. Vavilenkova
National Aviation University, 1, st. Kosmonavta Komarova

CONSTRUCTION OF THE TEXT'S CONTENT MODEL BASED ON LOGIC AND LINGUISTIC MODELS

The paper analyzes the shortcomings of existing models of text documents, it proposes an uniform content model of text that formed on the basis of synthesis of logic and linguistic models of its sentences. The work shows basic steps of the proposed algorithm for constructing content models of text.

Keywords – natural language, logic and linguistic model, content model, model of text, synthesis.

Problem statement

Despite almost a century of research in artificial intelligence, a computer still can not solve the problem of text information completely, because there are no adequate formal models of natural language objects and solving relevant problems includes informal, creative elements unique to human [1].

To describe the global meaning of the text we need to build a scheme which will provide quick analysis of surface structures and construction of relatively simple and rigid semantic configuration [2]. To build such a model we need to solve the problem of restoration of individual objects and their relationships, which were described in the text implicitly. The importance of solving this problem in information retrieval systems is the necessity to narrow the search, excluding documents that mention about useless objects to user and to protect ourselves from the fact that the user can request object by the words or phrase different that the author describes an event. That is why, the further study is directed to create a unified semantic configuration form of an arbitrary text document.

Analysis of recent research and publications

Today there are several models of representing texts that are used by modern systems of intelligent processing of text documents. Thus, a vector space model is used for the classification of electronic documents: each text is assigned with a point, each coordinate of which corresponds to a frequency of one of its words, so the texts with similar themes will have corresponding spatial vectors [3]. Another example of vector space model of text is Kohonen network used for quantization and the allocation of principal text's components [4]. Linguistic procedural model of text provides text analysis using various levels of processing (CPU, touch-sensitive register, long-term and short-term memory, etc.) [2].

Model of text under B. van Dijk and Kintsch is focused on the complexity of the description that involves a transition from understanding the components of sentences to sentences and text highest structures [5].

Traditionally, text is represented by histogram of homogeneous characteristics, such as the vector of frequency entries of keywords to the document. One of the most popular text models is "set of words" model, according to which the text is presented as a collection of words without regard to their order of use [6].

Konstantin Belousov in his work "Theory and Methodology polystrukturnoho synthesis text" [7] instead of the term "text model " introduces the term of "form of text" – a way of organizing linguistic substrate in a holistic space that has as its object a data structure where structure is a form of projection on the subject area.

However, none of the existing models do not allow to discover content relations in the text. This is because these models are based on statistical patterns of words of text base, but not on the logical relations between them.

The purpose of the article

The article is to describe the algorithm for constructing content model of text document on the basis of logic and linguistic models of natural language sentences. If we consider the text as a semantic structure, the main problem to be solved when building content model is how to segment the text and to classify quantum of information that form content. Indeed, various elementary snippets relate to situations, creating a hierarchy of semantic components. Thus, in the context of semantic space, text is considered to be a set of linearly arranged signs and a set of meanings [8].

The article is the result of the author's research in the text linguistics and formal semantics, combined with mathematical apparatus of first order predicate logic made it possible to formalize the description of the text documents.

Main material

Text is considered to be the universal mean of representation, stockpiling and transferring of knowledge, that is why the technologies of work with natural language texts have been always considered essential for artificial intelligence. Text is a set of interrelated one to another sentences that has a certain autonomy in relation to similar complexes; and substantial value that ensured unity communicative direction. The integrity of the text, converts it to a system in which elements depend on and provide each other (sentence, statement, paragraphs, sections). Therefore, the text can be formalized by creating its content model [9].

Logic and linguistic model (LLM) of text document is an abstract model that combines the basic properties of the text and its components, reflects the basic relationship between the structural components, it is an ordered array of four and set of logic and linguistic models of natural language sentences that compose the text [8].

Linguistic component of formal text description:

$$t =< CQ, F, B, A > , \quad (1)$$

where $t \in T$ – specific electronic text from the entire set of texts;

$CQ = \{cq_1, ..., cq_i, ..., cq_n\}$ – set of existing types of texts, $i = \overline{1,n}$, n – number of types;

$F = \{f_1, ..., f_j, ..., f_m\}$ – set of complex syntactic parts of the text, $j = \overline{1,m}$, m – number of complex syntactic parts;

B -text database consisting of a set of keywords and associated text proposals, and which can be represented as triples: $B =< K, SJ, D >$, K - a set of keywords; SJ – set of key phrases of the text; D – set of proposals;

$A = \{a_1, ..., a_k, ..., a_q\}$ – set of paragraphs of text, $k = \overline{1,q}$, q – number of paragraphs. Each paragraph in turn is described by triple: $a_k =< H, Y, R >$, $H = \{1, 2\}$ - a set of relations between sentences (parallel or chain); $Y = \{1, 2, 3, 4, 5\}$ – set of thematic progressions that taken in the paragraph $a_k \in A$; $R = \{1, 2, 3, 4, 5, 6, 7\}$ - set of thematic dominants in paragraph $a_k \in A$.

Let the natural language text is described by formulas within the set of T, then the logical formula $L(S)$ turns into identically true predicate in such an arbitrary interpretation, in which identically true predicate is converted to formulas within the set of T, that is $T \models L(S)$.

Semantic-syntactic part of the formal text description:

$$t' = \bigwedge_{g=1}^{N(t)} L(S_g) , \quad (2)$$

where $L(S_g)$ – logic and linguistic model [10] of natural language sentence S_g, $g = \overline{1, N(t)}$;

$N(t)$ – number of sentences in the text t.

Thus, the model (1) – (2) contains comprehensive information about the text and relations in it. Construction of such a logic and linguistic model for any type of text takes you to the analysis of textual information, comparison of texts by content, searching for contradictions and coincidences.

An algorithm for constructing of content text model using the logic and linguistic models involves the following steps:

1. Text segmentation. This step is responsible for the breakdown of the entire text information (electronic document) on several levels. Text segmentation is a function of the composition plan of the document, the main role is played by size of parts and content-factual information [11].

According to Halperin there are two types of text segmentation: three-pragmatic and context-variational [12]. The first type of segmentation is based on the quantitative parameter (the division into sections, parts, paragraphs, etc.). The second type of segmentation enables to establish the types of relationships between complex syntactic units, that is, to build a schematic structure of the analyzed text (in the text, there are three main parts: introduction, topic development and conclusion). Depending on the type of connected text, the number of above components could be greater so for specific text database there would be its own set of complex syntactic units F.

At this stage, the three-pragmatic segmentation is executed. It means a construction the structure of the document: the text is divided into sections, parts, paragraphs, and paragraphs in turn - on sentence. Text segmentation is purely technical step that does not take into account the syntactic and semantic relations.

Thus, after the phase of t-text segmentation there would be a set of paragraphs $A=\{a_1,...,a_k,...,a_q\}$, $k=\overline{1,q}$ received, where q - number of paragraphs and set of natural language sentences $S=\{S_1,...,S_g,...,S_{N(t)}\}$, where $g=\overline{1,N(t)}$.

2. Construction of logic and linguistic models of text's sentences. At this stage the t text is applied to method of automated construction of logic and linguistic models of natural language sentences, resulting that each sentence of the text is converted to a logical formula [13].

Implementation of this phase of algorithm ensures the formation of semantic-syntactic component of a formal text description (2), that is:

$$t'=\begin{cases} L(S_1)=\bigwedge\limits_{\mu=1}^{\nu(S_1)} L_\mu(S_1) \\ L(S_2)=\bigwedge\limits_{\mu=1}^{\nu(S_2)} L_\mu(S_2) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ L(S_g)=\bigwedge\limits_{\mu=1}^{\nu(S_g)} L_\mu(S_g) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ L(S_{N(t)})=\bigwedge\limits_{\mu=1}^{\nu(S_{N(t)})} L_\mu(S_{N(t)}) \end{cases}, \qquad (3)$$

where a simple predicate $L_\mu(S)$ for each natural language sentence is:

$$\begin{cases} p_{\mu[1]} & x_{\mu[1]} & c_{\mu[1]}(x_{\mu[1]}) & y_{\mu[1]} & c_{\mu[1]}(y_{\mu[1]}) & z_{\mu[1]} & c_{\mu[1]}(z_{\mu[1]}) & c_{\mu[1]}(p_{\mu[1]}) \\ p_{\mu[2]} & x_{\mu[2]} & c_{\mu[2]}(x_{\mu[2]}) & y_{\mu[2]} & c_{\mu[2]}(y_{\mu[2]}) & z_{\mu[2]} & c_{\mu[2]}(z_{\mu[2]}) & c_{\mu[2]}(p_{\mu[2]}) \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ p_{\mu[g]} & x_{\mu[g]} & c_{\mu[g]}(x_{\mu[g]}) & y_{\mu[g]} & c_{\mu[g]}(y_{\mu[g]}) & z_{\mu[g]} & c_{\mu[g]}(z_{\mu[g]}) & c_{\mu[g]}(p_{\mu[g]}) \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ p_{\mu[N(t)]} & x_{\mu[N(t)]} & c_{\mu[N(t)]}(x_{\mu[N(t)]}) & y_{\mu[N(t)]} & c_{\mu[N(t)]}(y_{\mu[N(t)]}) & z_{\mu[N(t)]} & c_{\mu[N(t)]}(z_{\mu[N(t)]}) & c_{\mu[N(t)]}(p_{\mu[N(t)]}). \end{cases} \qquad (4)$$

3. Synthesis of logic and linguistic models. In the third step of algorithm for constructing logic and linguistic model of the text document it is carried out a consolidation and replacement of structural components of logic and linguistic models (3) obtained in the previous step of the algorithm. This is done by identifying the ways of logical relation between natural language sentences.

On the basis of the principles of synthesis, logic and linguistic models (3) take the form as follow:

$$t' = \begin{cases} L^{(\gamma)}(S_1) = \bigwedge\limits_{\mu=1}^{\nu(S_1)} L^{(\gamma)}{}_{\mu}(S_1) \\ L^{(\gamma)}(S_2) = \bigwedge\limits_{\mu=1}^{\nu(S_2)} L^{(\gamma)}{}_{\mu}(S_2) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ L^{(\gamma)}(S_g) = \bigwedge\limits_{\mu=1}^{\nu(S_g)} L^{(\gamma)}{}_{\mu}(S_g) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ L^{(\gamma)}(S_{N(t)}) = \bigwedge\limits_{\mu=1}^{\nu(S_{N(t)})} L^{(\gamma)}{}_{\mu}(S_{N(t)}) \end{cases},$$

and, beside the replacements in formulas (4), every logic and linguistic model $L^{(\gamma)}(S_g)$ is attributed with vector of characteristics l_g, each of which corresponds to a specific component of logic and linguistic model of sentence S_e, related to the sentence of the text S_g:

$$\begin{cases} G_1(l_1): U \to u_k(S_e), e \neq 1, \\ G_2(l_2): U \to u_k(S_e), e \neq 2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ G_g(l_g): U \to u_k(S_e), g \neq e \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ G_{N(t)}(l_{N(t)}): U \to u_k(S_e), e \neq N(t). \end{cases} \quad (5)$$

4. Text database formation. Due to synthesis of logic and linguistic models there is a text database formed, which is a triple:

$$B = < K, SJ, D >,$$

where K - set of text keywords ;

SJ - set of key text phrases;

D - set of proposals.

1) Set of keywords in K text is formed from elements of vectors l_g, $g = \overline{1, N(t)}$, ie every element of logic and linguistic model that was replaced using of the principles of synthesis will be included to the set of keywords: $u_k(S_e) \subseteq K$, where $e = \overline{1, N(t)}$, when $e \neq g$.

2) Set of key text phrases SJ is formed from text phrases, that includes keywords from already created set of K. That is, if the word $H_r = u_k(S_e)$, $r = \overline{1, n}$ of sentence S_g forms a phrase with the word H_k, $k = \overline{1, n}$, $k \neq r$, the same sentence: $SJ_{rk} = H_r \cup H_k$, then $SJ_{rk} \in SJ$.

3) To form a set of proposals D it is applied an interpreter of productions that operates cyclically. The basic data for its work are obtained, due to the synthesis of logic and linguistic models, vectors of characteristics. Each element of the vector of characteristics l_g of natural language sentence S_g interprets relation of a simple predicate $L^{(\gamma)}{}_{\mu}(S_g)$ of sentence S_g and simple predicate $L^{(\gamma)}{}_{\mu}(S_e)$ of natural language sentence S_e, despite that $e \neq g$.

That is, as a result of work of the interpreter of productions it is searched such models $L^{(\gamma)}{}_{\mu}(S_e)$, the content of which precede or follow the models $L^{(\gamma)}{}_{\mu}(S_g)$:

$$L^{(\gamma)}{}_{\mu}(S_e) \to L^{(\gamma)}{}_{\mu}(S_g)$$

Text database construction stage enables to form meaningful links between sentences within text, no matter to which paragraph or complex syntactic part they belong.

5. Determination of each paragraph's characteristics. Determining the types of relations between sentences in each paragraph is one of the most important stages of linguistic analysis of the text.

Each paragraph $a_k \in A$ is characterized by triple of:

$$a_k = < H, Y, R >$$

1) Defining the first parameter $H = \{1, 2\}$ of the set of ***types of relations between sentences*** (1-chain or 2 -parallel).

2) Determination the ***type of thematic progression*** of the set $Y = \{1, 2, 3, 4, 5\}$, which is used in paragraph $a_k \in A$:

1. *Simple linear progression*. It is characterized by the consistent deployment of information when the theme of the previous sentence becomes the theme of the next sentence;

2. *Progression of cross-cutting theme*. It is characterized by one theme that is repeated in every sentence of the text;

3. *Progression with derived theme*. Each sentence of the text, without some kind of consistent elements of lemma-forming (the first type of progression) or cross-cutting themes (the second type) is used to express the general thematic direction of text;

4. *Progression with split theme*. The basis is a dual rhema, the components of which, while theme-forming, create the initial points for the development of separate thematic progressions;

5. *Progression with thematic jump*. It provides for a break in the chain of theme-rhema chain that you can restore from the context.

3) Determination the type of rhematic dominants of set $R = \{1, 2, 3, 4, 5, 6\}$ in paragraph $a_k \in A$. Depending on the lexico-grammatical representation the rhema may be of: subject, statal, dynamic, high quality, and combined.

Thus, due to setting properties of paragraphs we can identify thematic focus and logical means of building relations within them, so we can formalize the process of determining interphase links.

6. Formation of complex syntactic parts in text. The result of obtained relations between paragraphs in the previous step of the algorithm, text is divided into the complex syntactic parts(context-variational division):

$$F = \{ f_1, ..., f_j, ..., f_m \},$$

$$f_j \in F, \ j = \overline{1, m}.$$

Complex text parts are formed by using a set of text database D and the obtained values of set of rhemes R at the preliminary stage of the algorithm. After all, the purpose of rheme in text is not only in the representation of new content and updating communicative significance of information but also in organization the semantics of the text. Beyond the sentence's borders, the rheme comes into meaningful relations with rhemes of neighboring sentences, while it creates a rhematic dominant of text part, demonstrating their semantic unity.

Similarly to previous stage, we examine on means of cohesion not sentences but paragraphs. Based on their determination, there are following productions formed:

$$D_1 \to D_2,$$

where D_1 is the text's part from paragraph $a_k \in A$,that (by content) precedes or which implies text's part D_2 from paragraph $a_{k+1} \in A$, $k = \overline{1, q}$.

Based on these productions we can determine the leading relation in a complex syntactic part.

Determining the type c_i of connected text $t \in T$. Based on the previous defined types of meaningful relations between complex and simple syntactic structures, it has to be determined the type of

text $c_i \in C$, $C = \{c_1, ..., c_i, ..., c_n\}$ - set of existing types of texts, $i = \overline{1,n}$, n - the number of types (1-science, 2-journalistic, 3-art, 4-business).

Performing the above steps is an attempt to formalize semantic phenomena using linguistic facts. It relation to this, the study proposes a conceptual apparatus of semantic theory that provides explicit representation of the content of natural language sentences in the structure of a text document, consistent with the concepts of syntactic theory.

Conclusions

In grammatical terms, the connectivity of text is determined by the harmonization laws, rules of statement construction using morphological and syntactic means of language. In pragmatic terms, connectivity is induced by the general communicative function of the text, it is realized in subjective text organization, the system of spatial and temporal characteristics that permeate the text from beginning to end. [11] The study proposes logic and linguistic model of text document which takes into account both of these aspects.

The main step in algorithm of construction of meaningful model of text is a synthesis of logic and linguistic models, based on rules of construction and searching for elementary relations. The above relations have identical content and conclusively interpret natural language sentences of an arbitrary structure. Logic and linguistic model of text document is a kind of pattern, which an arbitrary text document can be reduced to.

1. Lyuher D.F. Artificial Intelligence: strategy solutions and methods slozhnыh problems / D.F. Lyuher - M .: OOO "Williams", 2005. - 4th edition Toe. - 864 p. 2. Elashkyna A. Krasnousova A. Maximov N. Rusin A. Post-rematycheskaya lynhvystycheskaya Model mashynnoy obrabotku text [Electronic resource] / Elashkyna A., A. Krasnousova, N. Maksimov, A. Rusin. - 2005. - Access for журн.:http://www.hr-portal.ru/article/tema-rematicheskaya-lingvisticheskaya-model-dlya-mashinnoy-obrabotki-tekstov#2.3.
3. A. Orlov Vektornaya model text [Electronic resource] / A. Orlov. - 2009. - Access for Zh.: http://neural.ru/dictionary/. 4. Kohonen T. Self-Organizing Maps / T. Kohonen. - New York: Springer, 2001. - Vol. 30 - 501 p. 5. Zvehyntsev V. A. New in zarubezhnoy lynhvystyke. Kohnytyvnыe aspects of language: Vol. 23 / Zvehyntsev V.A. - M .: IZ-in "Progress", 1988. - 320 p. 6. Sigachyov A.S. Model text in video recruitment chyslovыh pryznakov [Electronic resource] / A.S. Sigachyov. - Access for Zh .: http://it-claim.ru/Library/Books/ITS/wwwbook/IST7/ sigachov / Sigachov.htm. 7. Belousov K.I. Theory and Methodology polystrukturnoho synthesis text [Monohrafyya] / K.I. Belousov. - M .: Flynta: Nauka, 2009. - 216 p. 8. Vavilenkova A.I. Design of computer technology of linguistic analysis of electronic documents / A.I. Vavilenkova // Strategy of quality in industry and education: the ninth international conference, June 6-13, 2014. - Varna (Bulgaria), 2014. - P. 388 - 394. 9. Vavilenkova A.I. Methodological bases automatic analysis of logic and linguistic models /A.I. Vavilenkova // Mathematical Machines and Systems. - 2015. - № 1. - P. 65-71. 10. A.I. Vavilenkova The use of formal algorithms in structural linguistics /A.I. Vavilenkova // Proceedings of the National University "Lviv Polytechnic". - 2014. - № 699. - P. 265-272. 11. Golovkin S.H. Lynhvystycheskyy text analysis // S.H. Golovkin, S.N. Smol'nikov. - Vologda: Publishing Viru Centre, 2006. – 124p. 12. Halperin Y.R. As text object lynhvystycheskoho study. Ed. 5-Toe, stereotypnoe / Y.R. Halperin. - M: KomKnyha, 2007. – 144p. 13. A.I. Vavilenkova The use of formal algorithms in structural linguistics /A.I. Vavilenkova // Proceedings of the National University "Lviv Polytechnic". - 2014. - № 699. - P. 265-272.