УДК 004.934

A.A. Kharlamov, T.V. Yermolenko

# Text analysis: linguistics, Semantics, Pragmatics in the Cognitive Approach

Рассмотрены когнитивный подход к анализу лингвистической информации и процессы обработки информации разных лингвистических уровней: морфологического, синтаксического и др. В качестве примера приведена цепочка расширенных предикатных структур конкретного текста.

**Ключевые слова:** автоматическая обработка текстов, когнитивный подход, морфологическая обработка, синтаксическая обработка, семантическая обработка, прагматическая обработка, цепочка расширенных предикатных структур.

The cognitive approach to linguistic information analyzed by the human is considered. The processes of information processing are studied at various linguistic levels: morphological, lexical, syntactic and semantic levels for separate sentences, and finally, semantic and pragmatic levels for the text as a whole. As an example of the following processing, representation of the pragmatic level is identified as a chain of extended predicate structures of a particular text.

**Keywords:** automatic text processing, cognitive approach, morphological processing, lexical processing, syntactic processing, semantic processing, pragmatic processing, chain of extended predicate structures.

Розглянуто когнітивний підхід до аналізу лінгвістичної інформації та процеси обробки інформації різних лінгвістичних рівнів: морфологічного, лексичного та ін. Як приклад подано ланцюжок розширених предикатних структур конкретного тексту.

**Ключові слова:** автоматичне оброблення текстів, когнітивний підхід, морфологічне оброблення, синтаксичне оброблення, семантичне оброблення, прагматичне оброблення, ланцюжок розширених предикатних структур.

**Introduction.** Currently, two basic approaches prevail in the automatic semantic analysis of texts: linguistic and statistical ones. The first one provides a very detailed analysis of the meaning of the text sentences [1], and the second one makes it possible to create a semantic representation of the whole text [2]. They do not get along with each other; there are virtually no papers describing their joint application, which is explained by a significant difference of their implementation mechanisms. In the first case it is pure linguistics, and in the second case it is pure mathematics. However, their combined application could make it possible to obtain semantic representations of the whole text using fast algorithms of the statistical analysis with accuracy typical for the linguistic analysis.

There is a possibility of reconciliation of the linguistic and statistical approaches to the text analysis. For this we use the notions of information processing (including textual information) by human. To put it in a nutshell, the specific information processing in the human brain is reduced to its accumulation in the columns of the cerebral cortex of the brain [3], and its ranking in the hippocampus. The columns of the cortex are formed and stored dictionaries of event images (quasi-words from quasi-texts, including natural language texts) of various frequencies for various modalities. In the hippocampus ranking of these representations occurs, which characterizes the significance of these representations in individual situations (quasi-texts).

**Human linguistic information processing** considers the processing of text information at various levels (linguistic information – morphology, lexis, syntax, and supralinguistic information – semantics and pragmatics) in terms of the structural analysis, with natural transitions from one processing level to another one.

***Associative transformation.*** Cortical neurons collectively simulate multidimensional space and provide mapping of input sensor sequences in trajectories of this space [2].

Suppose we have an $n$-dimensional signal space $R^n$ and a unit hypercube $G^n \in R^n$ in it.

Using $G(n, N)$ let us denote a set of sequences of length $N$, the elements of which – points of the $R^n$ space – are vertices of the unit hypercube $G^n$. Here $G(1, N) \in R^n$ – the set of sequences of length $N$ ($N$ is an arbitrary natural number), the elements of which are binary numbers.

**Definition 1. The trajectory** is a sequence

$$\hat{A}: \hat{A} \in G(n, N)\ \forall n, N > 1. \qquad (1)$$

Indeed, if we consistently connect the points, which are the elements of the sequence $\hat{A}$, we obtain a trajectory in the $R^n$ space.

**Definition 2. *N*-termed fragment** is a fragment of length $n$ of the sequence $A \in G(1, N)$.

Let us introduce the transformation $F_n$ of the one-dimensional sequence in the trajectory $\hat{A}$ of the multidimensional $R^n$ space (2):

$$F_n : G(1,N) \to G(n,N+1-n),\ F_n(A) = \hat{A},\ (2)$$

where

$$A = (a(t) : a(t) \in \{0,1\})_{t=1}^{N},$$

$$\hat{A} = (\hat{a}(t) : \hat{a}(t) = (a(t+i-1), i = \overline{1,n}))_{t=1}^{N+1-n},$$

that is, $\hat{A}$ is a sequence of vectors $\hat{a}_n$ in the multidimensional space.

In the general case, the input sequence $A$ may contain similar $n$-termed fragments which results in self-intersection points of the trajectory.

The inverse transformation (2) is computed according to (3):

$$F_n^{-1} : G(n,N) \to G(1,N+1-n),\ F_n^{-1}(\hat{A}) = A,\quad (3)$$

Where $\hat{A} = (\hat{a}(t) : \hat{a}(t) = (a(t+i-1), i = \overline{1,n}))_{t=1}^{N}$, and

$$A = \left\{ a(i) : a(i) = \begin{cases} \hat{a}_1(i),\ 1 \le i \le N \\ \hat{a}_{i+1-N}(N),\ N < i < N+n \end{cases} \right\}_{i=1}^{N+n-1}.$$

***Formation of level-by-level dictionaries***. The memory mechanism that is sensitive to the number of passages of a given point in a given direction is a tool for analyzing the input sequence from the perspective of its repeating parts. As it is shown above, similar sequence fragments are mapped by the transformation $F_n$ into the same part of the trajectory $\hat{A}$ in the multidimensional $R^n$ space.

The dictionary forming is based on the analysis of multiple sequences $\{A_k\}$, in each of which, by superposition $H_hRMF_n$ (mapping $F_n$ the sequences of $\{A_k\}$ class, into the $n$-dimensional space, memorizing $M$ the number of passages by the trajectory of a particular point in the neuron memory, reading $R$ the contents of the memory of all neurons, and application of the threshold transformation $H_h$ to them) subsequences $\{B_j\} \subset A_k$ are identified that occur in it at least $h$ times (where $h$ is the threshold value of the threshold transformation $H_h$). Thus, the transformation $H_hRMF_n$ when interacting with the input set $\{A_k\}$ generates a dictionary $\{\hat{B}_j\}$ describing the trajectories corresponding to the subsequences $B_j$ of the input set in the $R^n$ space of a given dimensionality:

$$\left\{\hat{B}_j\right\} = H_h RMF_n\left(\{A_k\}\right). \quad (4)$$

Depending on the threshold $h$ value of the threshold transformation $H$, words of the dictionary $\hat{B}_j$ can be trees or graphs containing cycles.

***Formation of syntactic sequences***. The preformed dictionary can be used to detect old information ($\{\hat{B}_j\}$ dictionary words) in the new information stream (in the input sequence $\tilde{A}$ differing from the sequences of the set $\{A_k\}$ forming the dictionary). For this, the absorption of the $\hat{\tilde{A}}$ trajectory fragments of the input $\tilde{A}$ sequence is required that corresponds to the $\left\{\hat{B}_j\right\}$ dictionary words, as well as passing of new information (their links) regarding the dictionary.

To solve the problem of detection, the transformation $F_n^{-1}$ is modified to add detecting properties to it. Using the transformation $F_{n,C}^{-1}$ allows the formation of the so-called syntactic sequence or sequence of abbreviations $C$ characterizing the links of the $\{\hat{B}_j\}$ dictionary words in sequences of the set $\{A_k\}$. Let us denote by $\{B_j\}$ a set of subsequences corresponding to all chains of the $\hat{B}_j$ dictionary (4) words. Then:

$$F_{n,C}^{-1}(\hat{\tilde{A}}, \{\hat{B}\}) = C \quad (5)$$

$$C = c(t) : c(t) =$$

$$= \begin{cases} 0,\ \textit{если}\ \exists l,k : (\hat{\tilde{a}}(l),\ldots,\hat{\tilde{a}}(l+k)) \in \{\hat{B}\}, \\ \qquad\qquad l \le t \le l+k \qquad , t\text{-}1,..,N, \\ \tilde{a}(t),\ \text{or else} \end{cases}$$

$$\{C\} = F_{n,C}^{-1}\left(F_n(\tilde{A}), H_h RM\left(\left\{\hat{A}\right\}\right)\right) = \quad (6)$$

$$= F_{n,C}^{-1}\left(F_n(\tilde{A}), \left\{\hat{B}\right\}\right).$$

Thus, the mapping $F_{n,C}^{-1}$ allows elimination of some words contained in the dictionary $\{\hat{B}_j\}$ from the input sequence $\tilde{A}$. As a result, a structured approach to information processing is implemented: first elements of the structure are identified, and then links between them. The syntactic sequence $C$ containing only new information in regard to the dic-

tionary of this level becomes the input sequence for the next level of processing. At the next level, similarly to the level described above, the set of syntactic sequences $\{C\}$ forms the dictionary $\left\{\hat{D}\right\}$ and the set of syntactic sequences of the next level $\{E\}$. Thus, we have a standard two-level element of a multi-level hierarchical structure. Such processing with identification of level-by-level dictionaries occurs at all levels.

***Text analysis.*** In the text analysis at the stage of the morphological processing a dictionary of the first level is formed, $\{B_j\}_1$ – a dictionary of inflexions. Then the dictionary of the second level is formed, $\{B_k\}_2$ – a dictionary of stems. Next, the following dictionaries are formed: the dictionary of the third level $\{B_l\}_3$ – a dictionary of inflectional structures of syntactic groups, and the dictionary of the fourth level $\{B_m\}_4$ – a dictionary of pairwise occurrence of stems in the text. This co-occurrence is characterized by associations between these words, in other words, it means the semantic uncorrectness of the sentence ("Colorless green ideas sleep furiously").

Let us introduce the concept of the asterisk [4]. We will call a syntactic structure of the type:

$$d = <c_i <c_j>> = \cup_j <c_i\ c_j>, \qquad (7)$$

where $c_i$ is the dominant word, $<c_j>$ is a set of subordinate words, sematic features of the word $c_i$, an "asterisk".

**Statistical analysis of the text**

***Formation of the associative network of the whole text.*** The statistical analysis of the text is reduced to identification of the frequency pi of words in the text, and to identification of the pairwise occurrence pij of words in semantic fragments of the text. The pairwise occurrence characterizes the semantic co-occurrence of words in the language [5].

In simple cases of the text statistical analysis, to make the analysis more stable, and the results more interpretable, word forms of words are reduced to their radicals. At this a dictionary of stems $\{B_k\}_2$, and a dictionary of stems pairwise co-occurrence $\{B_m\}_4$ are formed. Thus identified stems serve further as the elements for constructing an associative (homogeneous semantic) network.

The associative (homogeneous semantic) network $N$ is a set of non-symmetrical pairs of notions (stems) $<c_i\ c_j>$, where $c_i$ and $c_j$ are notions (stems) connected with an associativity relation (co-occurrence in a text fragment, for example, in a sentence) $<c_i\ c_j> = B_i \in \{B_i\}_4$:

$$N = \cup_i <c_i\ c_j>. \qquad (8)$$

In this case, pairs of stems are linked through the same stems: $<c_1\ c_2>*<c_2\ c_3>$, where (*) means adjunction from the right. The result is a chain $<c_1\ c_2\ c_3>$, to which other pairs are further joined. At this branching and occurrences are possible, thus, actually a network is built.

If all pairs of words with the same first word are preliminarily grouped in an asterisk $d = <c_i <c_j>> = \cup_j <c_i\ c_j>$ (where $c_i$ is the dominant word, $<c_j>$ is a set of its semantic features), it can be said that the network can be built by groups of all asterisks:

$$N = \cup_i <c_i <c_j>>. \qquad (9)$$

***Notions reranking***. Elements of the semantic (associative) network $N = \cup_i <c_i <c_j>>$ and their links have numerical characteristics that reflect their relative weight in a given subject area – their semantic weight. To estimate the scale of semantic notions more accurately, weights of all related notions are used, i.e. weights of a "semantic constellation". As a result of the iterative reranking procedure, in each iteration notions associated with notions that have large weights, increase their own weight. Others lose it evenly:

$$w_i(t+1) = (\sum_{\substack{i \\ i \neq j}} w_i(t) w_{ij}) \sigma(\bar{E}). \qquad (10)$$

here $w_i(0) = p_i$, $w_{ij} = p_{ij}/p_j$ and $\sigma(\bar{E}) = 1/(1+e^{-k\bar{E}})$ is a function normalizing energies of all vertices of the network $E$ to the average value, where $p_i$ is the frequency of the $i$-th word in the text, $p_{ij}$ is co-occurrence frequency of the $i$-th and $j$-th word in fragments of the text (sentences). The resulting numerical characteristic of the words – their semantic weight – characterizes the degree of their significance (importance) in the text.

**Full linguistic analysis of the text sentences**

In full linguistic processing at the graphemic level of the analysis the text is segmented into words

and sentences, at the morphological level all the morphological information about words $\{B_j\}_1=\{m_j\}$ is identified, and at the syntactic level – the information about the links of words in groups and between groups $\{B_k\}_1=\{r_k\}$, where $r_k$ is a predicative link of the subject with the main object, and a $r_k \mid k>1$ are all other types of links. The structures of the syntactic level fall within the dictionary of templates for minimal structural patterns of the sentence and the dictionary of the verb valencies [6].

In the case of full linguistic processing for each simple sentence, an extended-predicate structure can be built, which after some transformations also reduces to an asterisk $d=\cup_j <c_i\, r_k\, c_j>$, where $c_i$ is a subject, $r_1$ – predicate, and $c_j$ – it actants. In the asterisk built from the extended predicate structure, the pair <dominant word, subordinate word> is complemented with a link between them marked with one of the $k$ types of the relation "predicate-actant" [7].

**Integration of the approaches. Semantic and pragmatic analysis of the whole text**

***Semantic analysis of the whole text.*** If an extended predicate structure of the sentence is identified using the full linguistic analysis of the sentence, then brought to the form of an asterisk, and then a semantic network is built using these asterisks, and its vertices are reranked, then a network is obtained in which associative links are replaced by the corresponding types of links. In this case, unlike an asterisk with simple associative links, in an asterisk built from the extended predicate structure, instead of pairs of notions triples $<c_i\, r_i\, c_j>$ are used, where between a pair of notions there is a link marked with one of the relation types.

***Formation of the text summary***. Next, let us consider what can be done with the text, and the inhomogeneous semantic network obtained from it. Since notions – vertices of the semantic network for a specific text – are ranked by their semantic weights in the analysis, we can use this to identify the portion of the sentences most significant for the text. We can calculate weight characteristics of the text sentences as a sum of weights for notions included in the sentence. Further, we can remove sentences, weights of which exceed a predetermined threshold. We will obtain a quasi-summary of the text. The cohesion of the text may be broken, but sentences contained in it will bear the meaning of the text.

***Formation of asterisk chains***. Separate sentences of the quasi-summary and the corresponding extended predicate structures describe separate fragments of the situation. The extended predicate structure has a corresponding (after the above transformation) asterisk $d=\cup_j <c_i\, r_i\, c_j>$. Then a chain of extended predicate structures contains the meaning of the quasi-summary:

$$D=(d_i \mid i=\overline{1,N}). \qquad (11)$$

where $N$ is the number of sentences in the quasi-summary.

**Example of the pragmatic analysis of the text**

Let us consider an example of pragmatic analysis of the text involving the described above mechanisms that allows identification of predicate structure chains for sentences of a text essential for representation of the text meaning. To simplify the interpretation of the chain, only most important parts will be taken from the extended predicate structures: (subject-predicate-main object).

As an example of an extended predicate structure of a sentence we take a sentence from T.I. Trofimova's textbook "Physics course", Moscow, "High School", 2001:

"Mechanics is a branch of physics that studies laws of mechanical motion and reasons that cause or change this motion".

We will not describe in great depth the details of the linguistic mechanism for extraction of the extended sentence predicate structure. Let us show the final result. The only remark is as follows: the sentence is broken down into simple components "Mechanics is a branch of physics" and "Mechanics studies laws of mechanical motion and reasons that cause or change this motion".

In the first part the extended predicate structure is very simple: "Mechanics (subject) – is included in (predicate) – physics (main object)".

The frequency of occurrence, the frequency of co-occurence are counted for the stems. After the formation of a semantic network, the frequencies of stems occurrence are converted into their semantic weights that allows calculation of the semantic weight of the sentences.

If sentences with weights less than the predetermined threshold value are removed from the text, what remains is the quasi-summary of the text, the fragment of which is shown in Table 1:

**Table 1**. Quasy-summary of the text (fragment)

| | Sentences of quasi-summary | Semantic weight |
|---|---|---|
| 1 | Newton's first law: every material point (body) persists in its state of being at rest or of moving uniformly straight forward, except insofar as it is compel to change its state by force impressed. | 99 |
| 2 | Newton's first law is true for all reference frames, and the frames with respect to which it is true are called inertial reference frames. | 97 |
| 3 | An inertial reference frame is a reference system with respect to which a material point, free from external forces either remains at rest or moves uniformly and in straight line | 99 |

The sentences of the quasi-summary reveal their extended predicate structures, which form the very chains (see Table 2) that characterize the pragmatics of the text. For the purposes of simplicity, below is a chain of only an essential part of the predicate structures (subject-predicate-main object). The other members of the extended predicate structures are omitted.

**Table 2**. The chain of predicate structures of the text (fragment)

| | Subject | Predicat | Predicat |
|---|---|---|---|
| 1 | Point | persists | state |
| 2 | Force | compel<br>change | it<br>state |
| 3 | Law | is true | NUL |
| 4 | Frames | are called | NUL |
| 5 | Frame | Is | NUL |
| 6 | Point | remains at rest, moves | NUL |

## Conclusion

This paper describes an approach that combines statistical and linguistic methods of text analysis, and semantic and pragmatic processing of texts using the proposed approach is demonstrated on specific examples. Combined application of the fast statistical algorithms of text processing, as well as linguistic algorithms and knowledge bases in the form of dictionaries of valencies make it possible to obtain the semantic representations of the whole text with accuracy typical for the linguistic approach. The understanding of the text pragmatics proposed in the paper is not, in general, universally accepted. However, such representation is sufficiently constructive to implement real mechanisms for the automatic text processing.

The works was performed within the research "Study of the mechanism of associative links in human verbal and cogitative activity using the method of neural network modeling in the analysis of textual information" (with financial support from the Russian Foundation for Basic Research, grant 14-06-00363).

1. *Leontyeva N.N.* Avtomaticheskoe ponimanie tekstov. Sistemy, modeli, resursy – M.: Academia, 2006.
2. *Kharlamov A.A.* Nejrosetevaya tekhnologiya predstavleniya i obrabotki informatsii (estestvennoe predstavlenie znanij). – M.: Radiotekhnika, 2006.
3. *Kharlamov A.A.*, *Raevsky V.V.* Networks constructed of neuroid elements capable of temporal summation of signals. / Neural Information Processing: Research and Development / Ed. by Jagath C. Rajapakse, Lipo Wang. – Springer-Verlag, May, 2004. – ISBN 3-540-21123-3. P. 56–76.
4. *Kharlamov A.A.*, *Raevsky V.V.* Perestrojka modeli mira, formiruemoj na materiale analiza tekstovoj informatsii s ispolzovaniem iskusstvennykh nejronnykh setej, v usloviyakh dinamiki vneshnej sredy // Rechevye tekhnologii. – 2006. – N 8. – P. 27–35.
5. *Rakhilina E.V.* Kognitivnyj analiz predmetnykh imen: semantika i sochetaemost. – M.: Russkie slovari, 2000.
6. *Dorokhina G.V.*, *Gnitko D.S.* Avtomaticheskoe vydelenie sintaksicheski svyazannykh slov prostogo rasprostranennogo neoslozhnennogo predlozheniya. Sovremennaya informatsionnaya Ukraina: informatika, ekonomika, filosofiya // Proc. of the Conf., May 12–13, 2011, Donetsk. – 2011. – **1**. – P. 34–38.
7. *Kharlamov A.A.*, *Ermolenko T.V.*, *Zhonin A.A.* The Understanding as Interpretation of Predicative Structure Strings of Main Text Sentences as Result of Pragmatic Analysis (Combination of Linguistic and Statistic Approach) / 15th Int. Conf. «Speech and Computer SPECOM'2013». – Springer, 2013.

*E-mail*: *kharlamov@analyst.ru*, *naturewild*71*@gmail.com*
© А.А. Харламов, Т.В. Ермоленко, 2015