The Patterns of Formalization of Nature-Language Messages in IT Security Monitoring Systems in Open Computer Networks

Victoria Korzhuk
St. Petersburg University of Information Technologies, Mechanics and Optics.
Kronverkskiy prospect 49, Russia

I. Introduction

In terms of social transformation taking place in the world it is necessary to supervise restlessly different information events. Integration of global computer networks to many fields of human activity causes emerging of IT resources that describe political, social and economic news and innovations. Messages of bloggers, agencies and data portals timeline commentators, Live Journal users contain information about attitude to developments in public life. In result the problem of automated data processing arises, and its purpose is to determine and analyze political, social and economic range of views.

Current easiness of using IT space granted by global computer networks provides a problem of ensuring IT security for objects in political, socio-economic, defense and cultural sphere of activity. Also specific damage of economic entity is caused by frequent using of different Internet resources for various PR-actions and IT-campaign that are created to solve political, economic and ideological questions so an analysis of huge amount of texts and documents for external and internal source of IT threat detection is necessary. However difficulties connected with using methods, that allow to identify the structure and the meaning of working nature-language messages in auto mode, lead to process this messages manually. But in addition high degree of integration and using PC along with implementation of IT technologies allows to develop and realize relatively advanced but more efficient methods and algorithms of semistructured data computation in IS [1].

II. The patterns of formalization of nature-language messages

Generally analytical patterns are highly tailored and too complex for adaptation to the concrete types of task of processing text information open computer networks.

To improve the quality of processing nature-language documents in the data domain of detecting information threats it is necessary to solve the problem connected with formalization of semantic component of text information in the messages.

One pattern that can be used for relatively short text messages processing is a semantic pattern of natural language proposed by Professor V. A. Tuzov of St. Petersburg State University [2]. It consist of 3 levels: morphological level, semantic-syntactic and semantic levels (Fig.1).

$$M=<W,Se,K>, \qquad (1)$$

where W – set of wordforms,

Se – set of semantic templates,

K – set of classes.

The feature of Pr. Tuzov nature-language pattern is united semantic-syntactic level. On this basis every word has morphological and semantic-syntactic characteristics which are the foundation for semantic predicate.

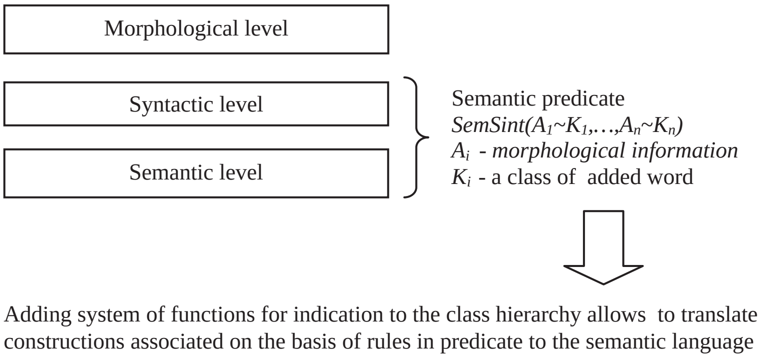
Morphological level
Syntactic level
Semantic level
Semantic predicate
$SemSint(A_1 \sim K_1,\ldots,A_n \sim K_n)$
A_i - morphological information
K_i - a class of added word
Adding system of functions for indication to the class hierarchy allows to translate constructions associated on the basis of rules in predicate to the semantic language

Fig.1. Semantic pattern of language by Pr. Tuzov

This pattern allows eliminating ambiguity of construction and reduces amount of noises in the document classification problem.

General wordform description template in the Pr. Tuzov's Dictionary is represented as

$G(Z1{:}!Nominative\{K_1\}_g,\ Z2{:}!Genitive\{K_2\}_g,\ Z3{:}!Dative\{K_3\}_g,\ Z4{:}!Absolutive\{K_4\}_g,\ Z5{:}!Instrumental\{K_5\}_g,\ Z6{:}!Prepositional\{K_6\}_g)$,

where $\{K_1\}_g \ldots \{K_6\}_g$ is a set of classes corresponding to a given wordform.

But Tuzov's semantic dictionary and Svedova's and Efremova's dictionaries that are used for the same tasks and also dictionary database of AOT and RCO companies are very different in structure, number of classes and the number of its constituent words. In result these products need additional adaptation for concrete text analyzing task connected with clarification of content and form (ex. arborescent or linear form) of wordform classificator.

The Pr. Tuzov's nature-language pattern suggests the possibility of analysis of every sentence of nature (Russian) language. Development of the using semantic data base occurred through the automated processing of different texts including literary texts. Due to random order of the words (ex. adjective can be separated from its noun by tokens so it located in another part of the sentence) it is necessary to make an exhaustion of all arguments to calculate the possibility of forming links for building nature-language structure of construction. On the other hand despite the support and development of this model there are certain troubles with computation of the result of sentence analysis because of emerging ambiguous wordforms, that influence on the construction of information objects. The high support cost are needed for using a pattern given here.

Adapted pattern which is designed to find concrete thematic information has fewer defects [3,4]. Similarly to the Tuzov's semantic pattern adapted pattern is divided into morphological, syntactic and semantic levels. Nevertheless semantic and syntactic levels are parted. Syntactic level contains information about links between words and semantic level defines the rules of the analysis, synthesis and processing of constructions.

$$M=<W, Si, Ks> \qquad (2)$$

where W – set of wordforms,

Si – set of syntactic templates, $Si \in Se$,

Ks – set of classes, $Ks \in K$

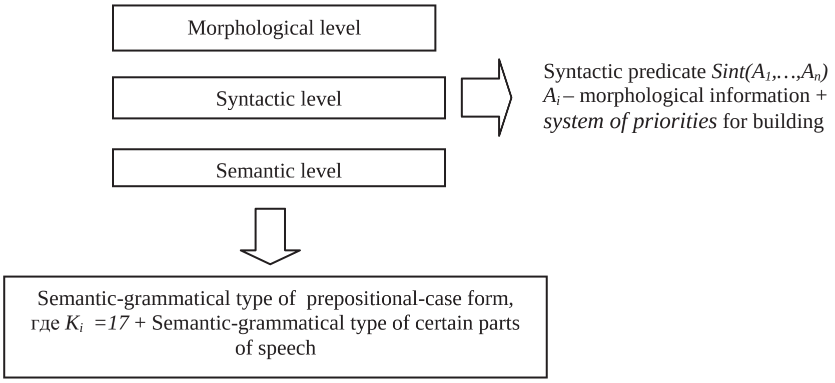
Morphological level
Syntactic level
Semantic level
Syntactic predicate $Sint(A_1,\ldots,A_n)$
A_i – morphological information +
system of priorities for building
Semantic-grammatical type of prepositional-case form,
где K_i =17 + Semantic-grammatical type of certain parts of speech

Fig.2. Adapted language pattern

Feature of this pattern is using of scalable predicates of wordform arguments information description of object-oriented dictionary data bases of natural language that allows to identify, to compare and to build control rules of processing at the level of links.

Scalable predicate is identical to semantic predicate of the previous model in composition. But here classes of identification sets which affect the type and semantic meaning of nature-language construction within the subject area are used instead of semantic class. Let us descry the construction and the features of it.

In our case analysis of stylistics in blog texts and time-lines of news agencies shows that long sentences are frequent in the works of Russian classics. Average length of such texts is about 10 words, and it is confirmed by statistic researches published on the dedicated to classical linguistics sites. Adjectives and qualifying nouns in the ablative and genitive, phrases which are identified with words “that”, “which”, “who” and some other and participles are not scattered on the message text but are close to the basic nouns that are forming construction. Assessment of the work of text information source of the Internet may be implemented through approaches based on the mistakes of the first and second kind. In this case dictionary databases adapt to the specific subject area. Limitations of subject area allow decreasing large number of ambiguous wordforms. Let us descry the simplified sentence convolution algorithm without focusing on the parts of speech and sentence, as numerals, conjunctions, particles, participles, gerunds and subordinate clauses.

Description of solutions for syntactic analyzer can be found at AOT company site (www.aot.ru). Principle of the algorithm is ordered sequential exhaustion method of about 40 rules.

But for text analysis in monitoring systems the most of the information is a noun. Its identification with followed accession of subordinate adjectives, adverbs, participles allows not spending resources on the calculation the type of formed constructions when the link forms. This algorithm uses the description of word-forms of parts of speech, based on a template containing syntactic information about potential links:

G(Z1:!Nominative, Z2:!Genitiv, Z3:!Dative, Z4:!Absolutive, Z5:!Instrumental, Z6: Prepositional).

Describing concrete wordform redundant links are removed. For example for the majority of the nouns syntactic pattern is

G(Z1:!Genitive).

Typical patterns of parts of speech and features of its using are show in [5]. The highest priority is given to the analysis of the possible formation of links between two nearest wordforms.

In simple extended sentence the following parts of speech: verbs, nouns, adjectives, adverbs may be contained (or not contained). The figure3 shows a sequence of steps of sentence convolution.

Simplified algorithm consists of the following steps:

1) Accession subordinate adjectives to nouns. Main information is taken from the morphological wordform descriptor. On the first viewing the proposals from left to right next in line adjectives and nouns that are consistent on cases, the gender and number, are searched. As an adjective may be the right from a noun, it requires a similar view from right to left, which makes an attempt to join the remaining adjectives were not included in the construction.

Due to space limitations, we will not dwell on individual cases where adjectives do not sequence on morphological information with their nouns, for example:

Tools and techniques - proven.

Such situations have a finite amount, and they are amenable to a fairly rigorous description and formalization.

2) Accession of prepositions to the nouns and adjectives structure. Feature of this step is that the preposition is always left from the noun construction. Main information for the implementation of the convolution is a syntactic preposition descriptor and morphological construction descriptor of the noun. The information about the preposition includes case and the using noun class.

3) Accession noun constructions to other objects is based on analysis of syntactic descriptor of left part and morphological and syntactic descriptor of right part and it is performed from left to right. Regardless of the descriptions the nouns object in the genitive case are attached to structures, standing on the left.

4) All completed constructions are substituted into the predicate of verb functions on the basis of their syntactic information.

5) Adverbs and assembled constructions not included in the descriptor verbs are attributed to it with its own semantic and grammatical type.

It should be noted that the Russian language is quite regular and exceptions to the rule amounts to not more than 10%.

Participial constructions, adverbial participle constructions, subordinate clauses beginning with words “*which*”, composite constructions like “*if ... then*” and embedded sentences should be separated before analysis. are exposed to the convolution algorithm, and then received constructions attached to the main proposal. All these constructions are subjected to convolution algorithm, and then received constructions are attached to the unitary clause.

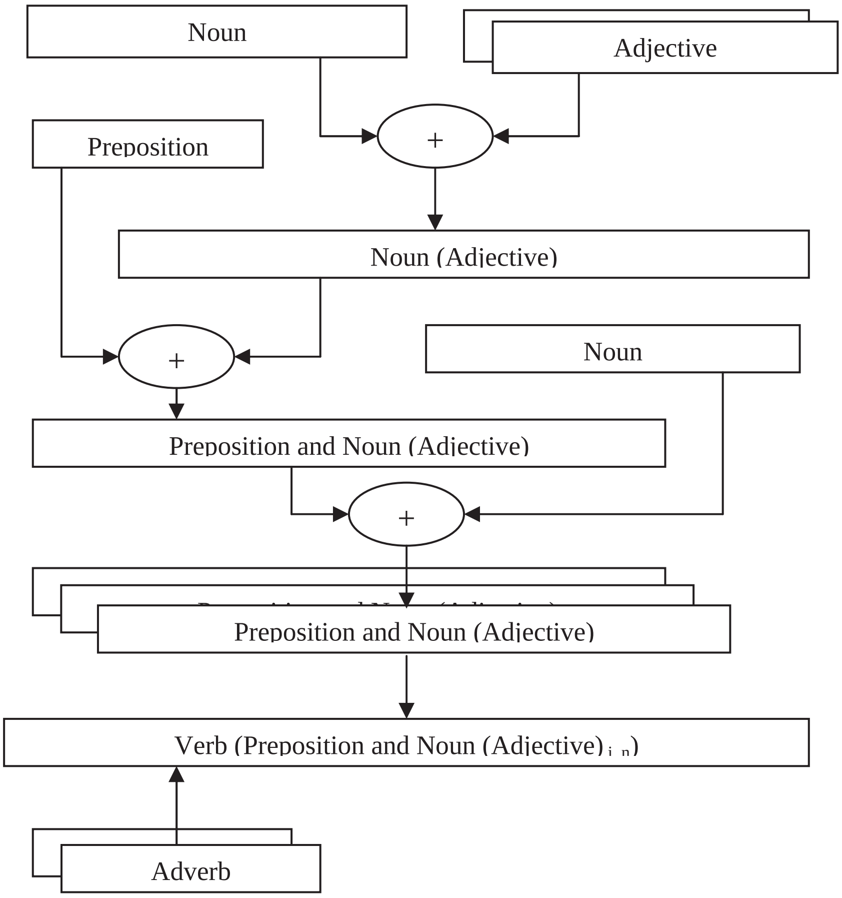
Noun
Adjective
+
Preposition
Noun (Adjective)
+
Noun
Preposition and Noun (Adjective)
+
Preposition and Noun (Adjective)
Verb (Preposition and Noun (Adjective)$_{i\ n}$)
Adverb

Fig.3. Simplified sentence convolution algorithm

Depending on the stylistic features of texts of the subject area and without grammatical errors parser produces 60% -80% of appropriate structures.

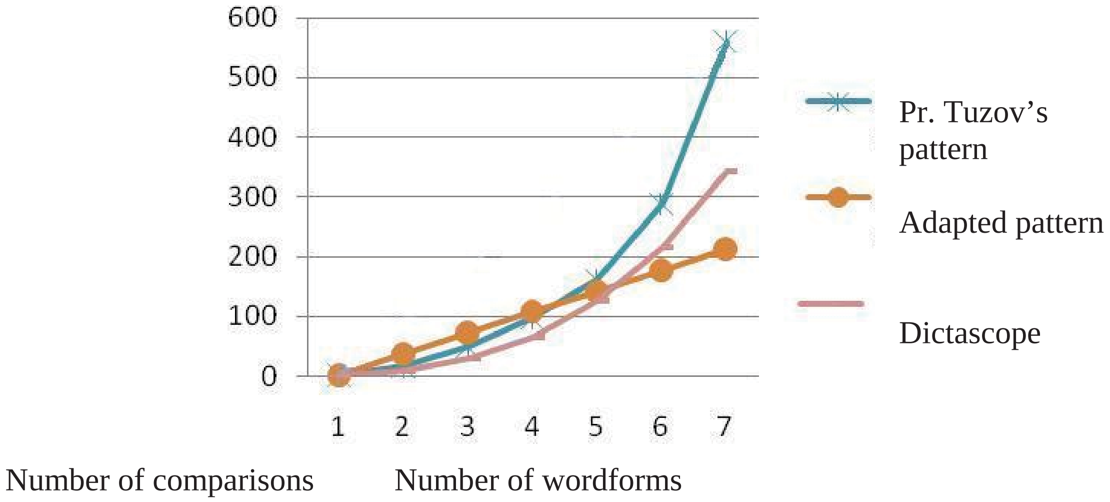

Fig.4. The dependence of the number of checks on the number of links word forms

Initial emergence of structure and superposition of semantic information on this structure allow to reduce the computational difficulties and to get rid of the exponential dependence of the number of analysis of links to the number of word forms of structures (Fig.4).

To realize analysis of textual information in the monitoring system an identification set $k_1 \ldots k_n$ should be initially configured in the database from a position of subject area of identifying text. To do this, analyzers from different vendors are used.

The processing of the sentences takes the form of functional record, containing the structure and links between its constructions.

$$F(f_i \to \{s\}_i) \tag{3}$$

where f_i is the words in the sentence each of that has its own set of links $\{s\}_I$ with other words.

Fig. 5 shows the links that form the other parts of speech relative to the prepositional-case forms of the noun. The vertices of this graph are a verb G, an adjective Pril, a preposition Predl, a noun S and an adverb Nar. Each arrow in the graph defined the set of questions that can be ask from different parts of speech to the prepositional-case forms of a noun or vice versa.

The first group is case questions group. It is almost unequivocally determined by the prepositional-case form and amenable to formalization at the level of syntactic template. The second group is a semantic questions group. For its formalization the classifier of nouns which are describing the semantic identity is requires.

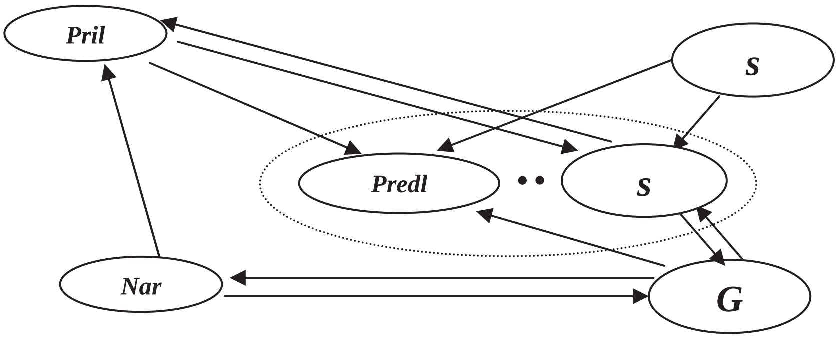

Fig.5. Links between the parts of speech regarding to prepositional-case forms of the noun

Texts run of the subject area through the parser allows to construct information structures and to carry out its statistical analysis for calculating the terms of the domain. Frequency of occurrence of the word, its context and constructions give information for building a classifier and for clarifying synonyms. Feature of this approach is that the basis of the classifier can be the third-party parser and the dictionary database.

In such way cited model of natural language uses scalable links predicate and its arguments contain information about the morphological characteristics and classes of adding words identifiers in wordforms description that can unify these descriptions and to simplify its structure.

Ensuring the economic, social and political security necessitates the audit of the information field and one of its tasks is to analyze the user's response to various events.

Modern processing system comments are aimed at getting an emotional assessment of messages. There are approaches based on statistical analysis in that messages wordforms are associated with semantic scales, such as *good-bad*. Each wordform of such scale is assigned a numeric value. Number of wordforms of the semantic scale in the commentaries allows to assess the general emotional state. However, in the debates and discussions a part of the identificators can not be related to the discussed events, but to other happens and objects. For example, you can find an anjective ”good” and an adjective ”bad” in a one part of sentence but associated with different nouns without any separating marks. In the case of a simple superposition of the good-bad scale given word forms characterizing the emotional assessment will affect each other. If you build the structure of nature-language construction it becomes apparent that the various information objects are defined.

Taking into account the style and the features of written comments in the Internet, consisting of the using of specific expressions and syntax errors in the construction of phrases and sentences, it should be noted that in the automatic mode it is not always possible to build an adequate structure of the analyzing message. In this case it is nessesary to use a universal approach to the construction of nature-language structures on the sintactic links level. In this problem information processing may be based on the calculation of the three kinds of elements: *objects*, *attributes and characteristics* and *actions*.

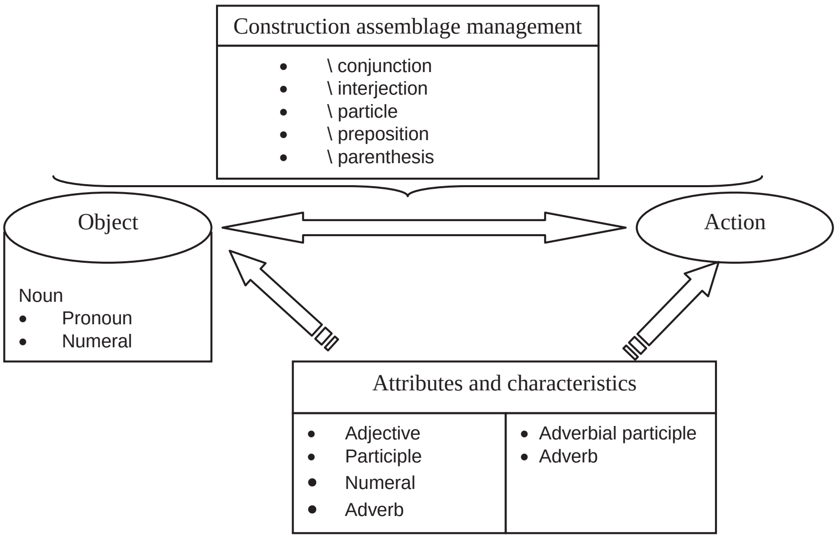

Fig.6. Universal structure of natural language representation

So the pattern that is the basis of obtained information structure can be described as:

$$M=<W,H> \tag{4}$$

where W is set of wordforms,

H is a set of attributes and characteristics

$H=\{O|D|C\}$

O is an object

D is an act

$C=\{C_o,C_d\}$ is the attributes and characteristics

Fig.6. shows the universal structure of the nature-language representation for the example of Russian language consisting of objects, actions, characteristics, and words which manage construction assemblage.

If we consider simple extended sentence in other natural language, it will be possible to compare the morphological identifiers according to the system described below.

1) Sentence objects are the nouns.

2) Action is a verb with its group which is determined by the sentence graph structure.

3.1) Characteristics of objects: adjectives, participles, adverbs, subject nouns.

3.2) Characteristics of action: adverbs, gerunds, adverbial participles.

4) Control words: simple and compound prepositions, punctuation.

Preparations phases for the simplest algorithm of creating the structure of sentence information objects based on morphological analysis consists of the following steps:

1) Searching of the sentence objects.
2) Searching of managing words.
3) Searching of the closest characteristics of the sentence objects.
4) Checking for the possibility of forming objects groups.
5) Action determination.
6) Searching of the action characteristics.

To implement the algorithm it is necessary to determine accurately the role of wordforms in a sentence and create a system of priorities for choosing a sequence of parts of speech. The problem solved with the help of this pattern is that messages text processing with the wrong syntax should be tried to get some related nature-language constructions on which can be define information objects, its characteristics, properties and actions. This pattern is a simplification of the previous ones described in this article and its advantage consists in fact that the proposed approach of creating a structure of universal constructions for most natural languages is quickly implementing without significant cost for the morphological and syntactic levels.

In the practical implementation this pattern is applied to the problems of monitoring and rating of statements of the events discussed in the Internet.

III. Conclusion

The approach to the selection of analytical patterns of representations of natural language in monitoring systems processing nature-language messages is based on providing the required characteristics (adequacy, completeness, accuracy) of the representation and reflection of textual information in databases and knowledge bases.

The detail level of properties calculating information depends on the structure of representation of the domain and subject area in a database of information systems.

References

[1]. Boyarsky K.K., Kanevsky E.A., Lezin G.V. Conceptual patterns of knowledge bases / / Scientific and Technical Bulletin SPbGITMO (TU). Issue 6. Information, computing and control systems. - St.: SPbGITMO (TU), 2002. P.57-62.

[2]. Tuzov V.A. Computer semantics of the Russian language. - St.: St Petersburg State University, 2004. - 400 pp.

[3]. Lebedev I.S. Way to formalize links in the construction of the text while creating a nature-language interface. / / Information and Control Systems, 2007, № 3. p. 23 - 26

[4]. Lebedev I.S .Building code templates for texts of the specification. / / Information Management Systems 2009, № 5. C. 39-43

[5]. Lebedev I.S. The construction of semantically related information objects of the text. / / Applied Science, 2007, № 5 (11). p. 83 – 89