OSTIS-2012
(Open Semantic Technologies for Intelligent Systems)

УДК 681.3.01

INTELLIGENT SYSTEMS FOR ENTITIES EXTRACTION BASED ON EXTENDED SEMANTIC NETWORKS

Kuznetsov I.P., Charnine M.M., Kozerenko E.B., Matskevich A.G., Nikolayev V.G., Somin N.V.

Institute of Informatics Problems of the Russian Academy of Sciences, Moscow, Russian Federation
igor-kuz@mtu-net.ru

Intelligent systems with linguistic processors extracting the entities and their links from natural language texts are considered. The conceptual model underlying the algorithmic developments is the extended semantic networks (ESN). The paper analyzes the use of the processors for text formalization in various subject fields: economy monitoring, criminal actions, mass media, terrorist activities (in Russian and English). Specific features of the texts are taken into account by linguistic knowledge of the processor: the system can be tuned to various subject areas.
Keywords: entities extraction, intelligent systems, knowledge mining, natural language, semantics.

INTRODUCTION

This work is dedicated to the questions of creating the engineering linguistic models of natural language for construction of linguistic processors for different classes of information systems and to description of the experience of the creation of linguistic ideas in the systems, which relate to the artificial intelligence research field. In the center of our attention are located the intellectual systems, developed on the basis of the apparatus of the extended semantic networks (ESN) [Kuznetsov, 1986], [Kuznetsov et al., 2009a,b]. We call them ESN-systems. These systems were created by the association of developers, including the authors of this article at the Institute of Informatics problems of the Russian Academy of Sciences during the period of two decades within the framework of research projects and applied systems, oriented at the concrete subject areas and customers.

Intellectual ESN- systems contain the developed bases of knowledge, in this case the knowledge is represented in the form of the records in the language of the extended semantic networks, called ESN - structures. Linguistic knowledge is, thus, a special case "of knowledge" and it is also represented in the form of the records in the language of the extended semantic networks. Basic structural element of the ESN is the named N-ary predicate, called "fragment". The whole set of language objects are given in the form of predicate-argument structures, in this case the mechanisms for presentation of embedded structures are supported, which gives very powerful presentation mechanisms for describing the objects of different language levels.

In the process of analysis and synthesis of natural language sentences the formal grammatical apparatus, similar to the dependency grammar, is used. With this approach the words and the constructions, which perform the role of predicates in the sentence, are the "support" elements, and the result of the analysis of a sentence must become one predicate, which corresponds to the predicate of the sentence (i.e. to basic verb in the tensed form or to another basic predicate expression) in question. Thus, in the process of analysis, in the first place, the processing is performed of the "action words" and the "relation words", i.e., of the verbs and other words, which have syntactic-semantic valences. An example of a "relation word" the word "father", "friend", and the like, i.e., in this case a "relation" is a word which assigns strong clearly expressed syntactical-semantic expectations. Semantic analysis in the engineering linguistic understanding is the process of translation of natural language expressions into "internal" structures of the knowledge base (KB) in our case these "internal" structures are the records in the ESN language. Thus, a KB structure is the code of sense in the intellectual information systems.

1. Basic aspects of semantic modelling

The procedure of conceptual-linguistic simulation on the basis of the ESN apparatus is based on the following principles: • a model must be "open" , i.e., support the effective mechanism of expansion and information update; • the model of the "sense" presentation should consider the facts of extralinguistic reality, which in the form of rules and relations compose a certain basic "world model" and the concrete models of subject areas; • the model should be practical, i.e., not overloaded by the detailed descriptions of connections and relations between the concepts in order to ensure the possibility of its realization, but at the same time, it should reflect the

relevant information for specific objectives.

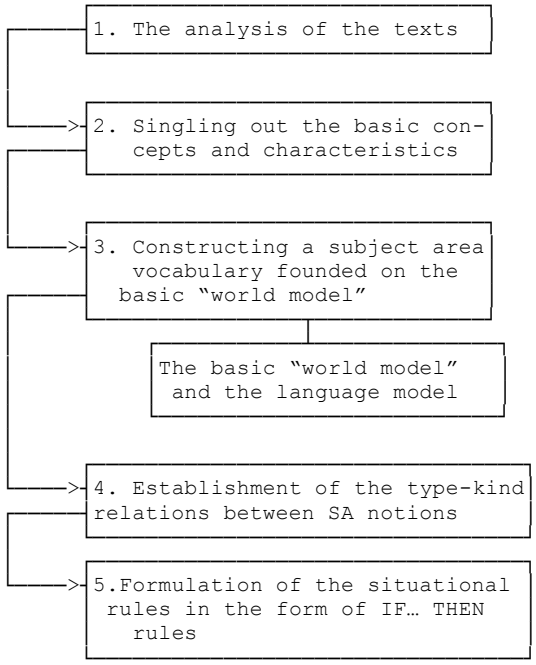

Figure 1 – The flowchart of conceptual linguistic modeling.

Realistic approach to the formulation of the problem dictates the need of limitation to a domain-oriented subset of a natural language. The essence of limitations consists in the following: - first, analyzed text materials contain expert knowledge from particular subject areas (we developed the systems for the subject areas for the diagnostics of the microcircuits production failures, forecast in the social sphere, criminology, and others); - in the second place, for the purposes of the maximally possible elimination of ambiguity, dictionary is built according to the modular principle: there is a certain most general common part (1-2 levels) completed by special dictionaries for each particular subject area. The proposed model of lexical semantics is based on the principle of the "nuclear" value realized in the context of this subject area with the subsequent inductive supplementation of other meanings (if they are actualized in the contexts in question). The taxonomy is also used which is realized in the form of the hierarchical trees of the word classes. The general "world model" of the system serves as the basis for the subject area models. The classes of words, are subdivided into concept/names, relations, actions, properties, characteristics of actions, time and place locatives. The most general notion is "concept", or universal class, which is subdivided into object, the situation, process and others. The words which relate to the classes of actions and relations, are represented as the semantic-syntactic frames, which determine the predicate-argument structures (government model). However, in the described approach (let us name it the ESN-approach) the range of argument values is substantially extended. This extension consists in the fact that in the role of arguments there can appear simple objects corresponding to the individual words, structural objects which present word combinations, phrases and clauses, and concept of "case" includes not only semantic, but also syntactic aspects. Approach, based on ESN allows to reflect the arbitrary level of the structures embedding it makes it possible to reflect the structural nature of lexical semantics, which in this model has a hierarchical network structure. Linguistic knowledge is represented in the system dictionary and the declarative modules of linguistic processor. In the ESN systems the function of dynamically formed semantic dictionary which is expanded automatically by the system in the course of concrete texts processing is also realized on the basis of initial linguistic information. In Fig. 2 the "internal" description of the verb in the semantic dictionary is represented. This dictionary is automatically generated by the ESN-systems DIES2, LOGOS-D, IKS in the course of natural language texts processing.

{(ВЫРАБАТЫВА895__)(DICSEM)
COORD(PROGNOZ1,RUS,ВЫРАБАТЫВА895__,S
50_31_51_20,%) SUB(UNIV,0+) SUB(UNIV,1+)
SUB(UNIV,2+)
ВЫРАБАТЫВ(0-,1-,2-/3+) INFI(3-)
ПРИДЕТСЯ(3-) ПРИДЕТСЯ(3-/4+) FUT1(4-)
SUB(СРЕД,5+)

Figure 2 - An example of the presentation of the verb *vyrobatyvat'* - "to manufacture" in the semantic dictionary.

2. The linguistic processor for entities extraction

The linguistic processor (LP) is realized by means of the language DEKL and is controlled by the linguistic knowledge (LK) in the form of object dictionaries, means of parametric tuning, and also the rules of extracting objects and connections [Kuznetsov, Kozerenko 2003]. With the aid of LK the tuning of LP to the appropriate categories of users and text corpora is accomplished. Concrete realization appears as a result. Thus, the paper deals with the means of constructing a class of processors with powerful mechanisms for their tuning and updating. Further development of such processors (LP) is connected with the development of implicit information, which we will consider in a narrow plan, i.e. as the addition of the structures of knowledge by the new information, which is absent or assigned implicitly. In this article the procedure of this development is proposed, which consists in the use of LP for mapping NL texts onto the structures of knowledge (ESN) and the use of the means of logical-analytical processing (productions of the language DEKL) for the creation of new information.

Advantages and deficiencies of the proposed procedure will be examined on specific objectives from the area of "criminology", that is the role functions establishment for the persons (participants) on the basis of the acts performed by them or due to the participation in some specific events. We consider the problem of assignment of properties to the persons (basing on their participation in the acts of different kinds) - "the suffered", "the

suspect" and others, if an explicit description of such properties is absent from the text. For example, if it is said in the text "suffered Ivanov I.I.", then another task appears, i.e. extraction of some property in the process of linguistic analysis and forming of the corresponding fragments in the knowledge structure. In this article the discussion will deal with LP, customized for the Russian language texts (NL), although the possibilities of LP are wider. There is a sufficient test of tuning LP to the English language texts.

3. The design method

The task of the role functions establishment for the information objects is a special case of the more general task, connected with the estimation of objects according to their descriptions in the NL texts, for example, with the estimation of the stability of enterprise (according to the information from the Internet), by featuring political figures (positive or negative depending on the statements in the press), by the estimation of the role functions quality of product (basing on the statements of users) and so forth. Quite frequently, it is not said directly whether something is bad, or good. As a rule, in NL texts the events are described, the situations, in which one or other information object participated. On the basis of them the estimation is done, which is often represented in the form of a new (generated) property of object.

For the solution of this problem different methods are used [Banko et al., 2007], [Clark et al., 2007], [Gildea, Palmer, 2002], [Pasca, Van Durme, 2007], [Punyakanok et al., 2008], [Cunningham, 2005], [Han, 2004], [Han, Kamber, 2006] . The most common one is the method of the new properties of objects development by using the syntactical-semantic forms. For example:

<what-medicine> caused allergy in <who-human organism>…,

< what-medicine > has side effects … <who-person> made scandal… and so forth.

The application of such forms to the NL texts consists in the search for "estimating" or "characterizing" words (of type "scandal") or for word combinations of the type "caused allergy" ("it can cause allergy"), it "has side effects" ("side-line actions"), "to make scandal" ("to brawl")… And then the environment is analyzed, i.e., the words, which stand to the left and to the right, their semantic classes (objects are recognized by them) and case forms. Estimations of information objects as a result are given. By the first two forms the "quality of medicines" is estimated, while by the latter it is recognized that a man performed "hooligan actions" or that he is "suspected". It is known that in NL many versions are possible for expressing the same idea - with the aid of different syntactic constructions, verbal groups, forms and so forth.

Therefore the number of estimating word combinations will be sufficiently large. Moreover, the application of such forms requires different forms of analysis - morphological (in order to reduce different word forms to one form), syntactic (the trees are built of the selection of sentences in order to isolate the connected components and to find place for the estimated words) and semantic (in order to extract the objects, which are evaluated). The use of syntactical-semantic forms is connected with certain difficulties caused by special NL features: by the presence in texts of participial, verbal-adverbial constructions, different explanations, facultative components (time, place, purpose), anaphoric references and other language structures. As a result, information objects are frequently disconnected from the estimated words. Hence - the significant losses, which influence the quality of estimation.

Example 1 *(the text is taken from the summaries of incidents of the City Office of Home Affairs, Moscow): … Gorelov Peter Sergeyevich, 01.03.76 yr/bir, liv: c. Moscow, st. Young Leninists, h.71-6-12, does not work, 01.02.1998 yr. at 4.30 in his house out of hooligan motives in the state of alcoholic intoxication made scandal and broke window glass in the apartment of Litvinova Galina Ivanovna, 20.07.1961 yr/bir,…*

In this example the estimating (characteristic) words are "made scandal" and "broke the window glass", they are located at a significant distance from the estimated person - "Gorelov Peter Sergeyevich". This limits the possibilities of applying the forms. It is required that the initial extraction of components, which must not be considered in the forms: the years of birth, addresses, specific properties ("he does not work", "in the state of alcoholic intoxication"), time, place and others, which requires sufficiently deep text analysis with the extraction of objects, their properties and attributes. In connection with the aforesaid, another more promising method is represented - when evaluation is accomplished at the level of knowledge structures. For their construction the objective-oriented LP is used producing the structures of knowledge in which the objects are directly connected with the events and the actions and excluding the above mentioned losses. For the development of implicit information (role functions of objects) the rules of the DEKL language are used which analyze the structures of knowledge (ESN) and form new properties of objects. In this case the structure of knowledge does not change, but it is only supplemented by new (useful) fragments.

4. Presentations of the meaningful portraits of documents

Within the framework of the proposed procedure the development of the role functions of objects (implicit information) is achieved at the level of the structures of knowledge, called the meaningful portraits of documents (SS-documents). Let us examine how such structures appear in the ESN formalism .

Example 2 (translation of the Russian text given below). *Text N22 is taken from the summaries of incidents of the City Office of Home Affairs, Moscow*:

01.02.98 yr. 16-30 to the Home Office applied citizen Mitrofanov Victor Mikhaylovich, 1955 yr. bir., liv.: Bohr Highway 38-211, n/w. he stated that 01.02.98 yr. at 10-00 in house 3 at St. Fedosino the unknowns being found in the drunk state made scandal, they expressed themselves by unquotable swearing, they set dog. As a result of what Mitrofanov applied to trauma care center, where the diagnosis was set: the bite of foot.

The objective-oriented LP performs the deep analysis of the text and automatically builds its meaningful portrait (SS- document, transliterated):

DOC_(22, "1-02-98", "SUMMARY; " /0+) 0 (RUS)
OVD_(OVD/1+)
FIO(MITROFANOV], VICTOR, MIKHAYLOVICH, 1955/2+) UNEMPLOYED (2-/3+) 3- (22, PROP)
ADR_(Borovskiy, Sh., 38,211/4+)
PROZH. (it is 2nd, 4)
ADR_(UL, FEDOSINO, HOUSE, 3/5+)
FIO (" " , " " , " " , NESKOLKO/6+)
UNKNOWN (6)
DRUNK (6-/7+) 7 (2, PROP_)
SCANDAL (6, PYANYY/8+)
IS EIGHTH (22, ACT_)
TO REPORT (IT IS 2ND, 8-/9+) 9 (22, ACT_)
DATA_(1998,02, ~01, " 10-00" /10+)
When (9, 10)
TO TURN (1, GR- N, 2-/11+) 11- (22, ACT_)
DATA_(1998,02, ~01, " 16-30" /12+)
When (11-, 12-)
EXPRESS (6, UNQUOTABLE, [BRAN]/13+) 13- (22, ACT_)
TO SET (6, [SOBAKA]/14+) 14 (0, ACT_)
TO TURN (IT IS 2ND, IN, [TRAVMPUNKT]/14+) 14 (0, ACT_)
TO PLACE (DIAGNOSIS, BITE, [NOGA]/16+) 16 (0, ACT_)
PREDL_(22,11-, 4, 3-, 9, 13-, 14-/17+) 17- (2,15,341)
PREDL_(22,15-, 16-/18+) 18- (6,342,448)

A meaningful portrait consists of the elementary fragments, arguments of which are words in the normal form (necessarily for the search and processing). Each elementary fragment has its unique code, which is written in the form of the number with the sign + and is separated by a slash line. For example, in the fragment OVD_(OVD/1+) the sign 1+ is its code (but 1 is the reference to it). Fragments DOK_(22, "1-02-98.TXT", "SUMMARY; " /0+) 0 (RUS) indicate that the meaningful portrait is built on the basis of the Russian-language text of document with number 22 of the file of 1-02-98.TXT", which was processed as the summary of the incidents (linguistic knowledge depend on this). The following fragments present police department OVD_(… /1+), person's surname, name and patronymic FIO (… /2+), person's specific property UNEMPLOYED (2-/3+), address ADR_ (… /4+) and so forth; the signs 2+, it is 2nd, 3+, 3-,… are the codes of the fragments, with the aid of which their connections and relations are assigned. For example, the fragment PROZH (live) (it is 2nd, 4) represents the relation that the person (represented as FIO with code 2+) lives at the address (fragment [ADR]_ with code 4+). Actions are represented in the form of fragments of the type SCANDAL (6, PYANYY/8+) it is 8 (22, ACT_), where it is represented that "person (FIO with code 6+), being drunk, made scandal". With the aid of it is the fragment 8_(22, ACT_) indicates that the first fragment is SCANDAL (…./8+) presents the action and relates to the document with the number 22. A similar role is played by the fragments of the type 3-(22, PROP_), by which the properties are noted. The codes of fragments also serve for the idea of time, scene of action and cases, when one action is included in the composition of another. For example, the fragment TO REPORT (it is 2nd, 8-/9+) represents that the person (code 2+) "reported" (code 9+) about the action (code 8+), i.e., about "made scandal". The following fragments DATA]_(… /10+) when (9, 10) represent the time (DATA_), which relates (when) to the action "to report". Special role is played by the fragments PREDL_(...), which correspond to the sentences. They are filled up with the words, which did not enter the information objects (in this example they are absent), or with the codes of objects themselves.

To these fragments the indicators of their position in the text are added. For example, the fragment PREDL_(22,11-, 3-, 9, 13-, 14-/17+) 17- (2,15,341) represents the fact that the objects with codes 11- (corresponding to the action "to turn"), 3- (corresponding to the property "unemployed") and others are located in the sentence, which begins from the 2nd line of the text of the document and they occupy the place from the 15-th to the 310-th byte. These means of positioning are necessary for the work of the reverse linguistic processor (LP).

Analyzing this example, it is possible to make the following conclusions: 1) In SS- document the estimating (characterizing) words occur either in one fragment with the object - SCANDAL (,…), or the next one, i.e., the codes of the actions, in which the object participates, are nearby in PREDL_(… 9, 13-, 14…). In this case the possibility of composite actions is considered. 2) On the actions, represented as SCANDAL (,…), it is possible to draw the conclusion that the discussion deals with "that suspected", and TO REPORT (,) - that the person is "suffered" or "the applicant". Such conclusions are easily arrived at with the aid of the rules IF… THEN (productions) of the language DEKL, which are the basis for the extraction of role functions. 3) The particular difficulties of dividing the text into the sentences occur (in the old version).

The reduction "of n/r" (with the point at the end) was not understood as the end of a sentence. 4) The linguistic processor (LP) correctly identified the

pronoun "*he*", and also it knew how to reveal the participation of the subject ("*unknowns*") by the actions "*to be evinced by unquotable swearing*" and "*to set dog*", which also characterize subject. At the same time the LP could not connect the action "*diagnosis was set*" with the person - "*Mitrofanov*…" (the code is 2-nd). In this case an example proved to be successful. Also the processor LP (with its linguistic knowledge - LK) was developed for the tasks of the criminal police, connected with different forms of the objective searches: the search for similar participants (addresses, and so forth), search according to the connections, precise search for objects, for the search by signs and other identifiers. In this case the analysis of some complex NL forms was not required, i.e. the cases of the enumeration of the objects participating in the uniform actions (they are described by one verb), the enumeration of the actions of one object and others in contrast to the aforesaid, with the extraction of role functions for each object the indication of its participation in each action is required. Hence it follows that with the use of the proposed procedure the more qualitative extraction of role functions is directly connected with the works on improvement of LP in the aspect of the development of objects and their actions. In many instances the numerous errors caused the inaccuracies in SS-document, e.g.: the absence of punctuation marks or their presence (where it was not required), the inappropriate reductions, gaps in the words and many others. The fact is that the documents, entering the summaries of incidents, are composed on the spot by people (militiamen) of different degree of literacy. Hence – the additional noise and loss. Thus, meaningful portraits are the collections of fragments of ESN which represent the sufficiently high level of formalization of NL texts and are convenient for the working - with the aid of the instrument means - DEKL. Besides LP which analyzes texts and builds SS-documents, there is a reverse linguistic processor (LP) which on the basis of the fragments of the SS-document generates the NL texts presented to the user.

5. The means for the establishment of the role functions

Within the framework of the proposed procedure (instead of the application of syntactical-semantic forms to the documents) the rules are used for logical conclusion and transformation of the knowledge structures - the SS- documents, in which there are no morphological features (of type who, whom,…), and the subjects and the objects are distinguished by their arrangement in the fragments of ESN, which present actions. The names of fragments present the nature of actions. Syntactical-semantic forms are transformed into the fragments of ESN which determine conversions and logical conclusion achieved by productions of language DEKL. Such fragments play the role of the logical-semantic shell, which determines conversions and logical conclusion on the basis of SS-documents. After filling of the shell by ontological-fragmental knowledge (OFK) which consist of the mentioned fragments (ESN), the program is formed, which accomplishes the development of role functions and completion of the SS-document by the appropriate fragments. With this approach it is possible to avoid many difficulties, connected with the design features of NL and the specific character of the use of syntactical-semantic forms. There are many versions of construction of the shells and representation of the corresponding knowledge which are distinguished according to the degree of their generality. Let us examine the version which is at present realized and verified.

Case 1. The role functions are determined by the names of actions. In this case for the extraction of objects (participants) which should be assigned properties (role functions), the fragments of the following form are used :

INTERPRET (MAN_2, FIO, " suffered") FORMA_CC (MAN_2, CLASS_D4, " ") CLASS_D4 (TO TURN, TO STATE, TO REPORT, TO PASS AWAY,…)

The first fragment INTERPRET (…) means that from the SS- document it is necessary to extract the fragments of the form FIO (…), that correspond to participants, and to analyze the possibility of assigning them the property "suffered". Such participants are conditionally designated as MAN_2. The second fragment FORMA_CC (…) specifies the conditions for assigning this property to MAN_2, determined by the constant CLASS_D4. In the third fragment CLASS_D3 (…) the words are given which present actions. It is represented that the words belong to the class CLASS_D3. If the participant occurs in one of the enumerated actions, then to this participant the property "suffered" is assigned. This participation is revealed via the analysis of the SS-document. If there is a fragment TO TURN (…, it is n-th,…) in it, the argument of which is the code FIO (… /N+), then the fragment N-("suffered") is added that represents the role function of the corresponding participant. Conformably for the SS- document represented in example 2 the analysis will occur as follows. Consecutively the extraction of fragments FIO (…) corresponding to the participants is performed. First FIO (MITROFANOV,… /2+) will be extracted . Its code is 2- is the argument of the fragment TO APPLY (1, GR- N, 2-/11+), that presents the action. In connection with this to SS- document the fragment 11-("suffered") will be added, which via the reverse LP will be transformed into the statement that "Mitrofanov Victor Mikhaylovich is a suffered person". These actions are realized within the framework of the logical-linguistic shell.

Case 2. Role functions are determined by the actions and elucidating words. For this the same fragments are used, as in the first case, but during the enumeration of the names of actions the additional fragments which present actions with the possible elucidating words, are introduced:

INTERPRET (MAN_1, FIO, "suspect")
FORMA_CC (MAN_1, CLASS_D3, " ")
FRAUD (USER, POKUPATEL/15+)
TO SET (DOG/16+)

TO BE EXPRESSED (UNQUOTABLE, SWEARING, OBSCENE,… /17+)

CLASS_D3 (IS DELAYED, TO BE SOUGHT,…, 15,16-, 17-)

The given fragments determine actions of the extraction of persons (MAN_1), by which the property of "suspect" is assigned. For this at the level of the knowledge structures their participation is analyzed in the actions "*is delayed*", "*to be sought*", and also in the composite actions: "*to set dog*", "*to be expressed unquotable*…", "*to be expressed by swearing…"* and others. In example 2 the code of fragment FIO (" " , " " , " " , FEW]6+), that represents the unknown persons is the argument of the fragment TO SET (6, DOG/14+), representing action *"to set"* with the elucidating word *"dog" – "sobaka"*. Therefore the fragment 6 is added ("*suspect*"), that represents that "*the unknown persons are suspected*", and through the reverse LP the explanation to this conclusion is offered, see below. A similar conclusion will be made on the basis of the fragment TO BE EXPRESSED (6, UNQUOTABLE, SWEAR/13+), but with other explanations.

Case 3. The actions determine the role functions of several persons. For this (additionally to the fragments INTERPRET) the fragments are added: CLASS_D1 (TO STRIKE, TO BEAT UP,…) FORMA_CC (MAN_1, CLASS_D1, MAN_2), where FORMA_CC (…) indicates the need of the search of two persons - "*suspect*" and "*suffered*" (MAN_1 and MAN_2), that participate in one action, which are mentioned in the fragment CLASS_D1 (…). For example, "*certain person struck another*…". In the appropriate fragment TO STRIKE (…) the code FIO (…) that corresponds to the first person will stand in front of the second. The given fragments of ESN compose the knowledge OFK which are constantly supplemented - due to the filling of classes by the new words-actions and with the elucidating words. The process of filling is sufficiently simple. If role function is not revealed, then it is necessary to look in the SS- document in which the action of one or another participant (by the text its role is easily determined) occurs. Further, the corresponding constants are located, by which are supplemented the classes of knowledge OFK. Subsequently it is intended to automate the process of completing the knowledge OFK as follows. In the text the words, which determine role functions, are noted. Further, in the formed SS-document the corresponding constants which supplement knowledge OFK are located.

6. Factors of processor quality

The quality of a linguistic processor is determined by a number of factors. First, the possibility for isolation of objects and connections. These are the types of objects being isolated, their quantity. The Semantix processor identifies up to 40 types of objects, including very complex ones, which correspond to actions and events. With an increase in the quantity appear the additional difficulties, connected with collisions of the extraction rules of: some rules can seize the words, which relate to other objects and those extracted by other rules. It becomes important to consider the order of the application of rules, including of the rules of identification. In the second place, an important factor is the selectivity of rules and procedures of the identification: the factor of the noise and losses. By noise we mean the presence of excessive words in the objects. Losses are the situations when an object is not revealed or revealed partially: in the text there are the words, which did not enter into the object. In the Semantix processor the rules are arranged in such a way that they ensure the high degree of selectivity and the minimization of noise and losses with the large number of the objects being selected.

The third factor is the possibility and the labor expense for tuning to a corpus of texts (for increasing the selectivity of rules for extraction of objects), and also tuning to the new objects. Due to the complexity of analysis this tuning should be achieved through the linguistic knowledge (LK). The latter should have all means for increasing the selectivity of rules and necessary conveniences in the plan of their creation and correction. Ideally, with the aid of LK the tuning to the special features of language as well as to the standard language forms should be ensured. The Semantix linguistic processor ensures the analysis of the Russian and English lanuage forms with the aid of the uniform language model.

The fourth factor is the speed of linguistic processor operation, i.e., the time of text analysis. The speed is determined by the design features of a processor (by means of search time decrease), and also by the number of objects being extracted. The application of rules of extraction is connected with the search for the necessary words, where sortings are required. The greater the number of objects and rules, the greater the time of analysis. In the Semantix processor there are different means of sorting time decrease. Besides the program there are also means of control by linguistic knowledge. It is indicated for each rule, what words should be searched for the initiation of the process of its application. The permissible contexts (to the left and to the right of revealed words) are assigned. These features ensure sufficiently high speed (fractions of a second for 1 KB of text) with a sufficiently large number of objects extracted. More than 40 different types of objects are supported by the Semantix processor. The subject areas represented in the text documents are as follows.

- Documents about terrorism in the Russian language. The analysis of the documents, in which the discussion deals with the terrorist acts and the groups. This feature supports the extraction of 40 types of objects, their connections and the degree of participation in the criminal actions.
- Documents about terrorists in the English language. The objects and links include persons (their family name, name, patronymic – FNP), posts, organizations, terrorist groups, instruments of crime,

time and place of events and so forth, and also connection with and participation in the actions.

- Summaries of incidents. Is ensured the extraction of figurants, their connections, organizations, dates, documents, numbers of bank accounts, details of weapons, etc. with the indication of their participation in particular criminal actions.
- Accusatory conclusions, information about the criminal cases. Objects are identified along the entire field of text. Their connections and criminal actions are revealed.
- Government communications, media issues. Persons, dates, organizations, positions and other significant information and also connections and participation in the actions are selected.
- Autobiographies in the Russian language. From the Russian language resumes all attributes of people, periods of time and place of their work, studies, language proficiency and so forth are extracted.
- Autobiographies in the English. From the English language resumes are all attributes of people, periods of time and place of their work, studies, language proficiency and so forth are extracted.
- Documents of media issues in English. From the English language texts the persons mentioned in media issues, positions, organizations, dates, terrorist and anti-terrorist groups, weapons, events, their time and place, different connections and other features are extracted.

In the processors of the Semantix, Lingua-Master, "Criminal" systems up to 40 types of objects are extracted with high accuracy and minimum noise. For example, the system "Criminal" was verified on about 500 thousand incidents from the summaries of Moscow Criminal Police Department, and on the basic objects showed the unique results: the coefficient of noise, i.e. excessive words in the objects) is not more than 1-2% and losses are not more than 1%. The Semantix Processor was fixed on a smaller quantity of documents dealing with the terrorist activity, and therefore there can be more noise and losses in it. But this can be quickly fixed. The fact is that to consider everything which can be encountered in the NL texts is impossible. Therefore, in the first place, the representative collections of test documents are extremely important, and in the second place, the means of fixing or tuning of linguistic processors are as follows: the employment of hybrid approaches comprising hand-made rules and statistical means for rapid correction and fine adjustment of linguistic knowledge.

In our systems there is an entire complex of such means which ensure rapid tuning to the applications (including the introduction of new objects and connections) taking into account the demands of customers. A sufficiently in-depth analysis of sentences is conducted with the development of verbal forms, and also with the identification of objects of the entire text. The analysis of the complex language structures is ensured: forms with verbal nouns, participial and adverbial constructions, coordinated terms, etc. is supported by the expert component. The Semantix processor can be used as a stand-alone (independent) module. At present the first release of the English language version of the object - oriented linguistic processor Semantix has been developed.

CONCLUSION

The objective-oriented linguistic processor for entities extraction can be used in different areas of application. It has a number of essential advantages, the main is that it outperforms the recently appeared systems by the number of the supported object types, there is an entire complex of logical and statistical means which ensure rapid tuning to the applications (including the introduction of new objects and connections) taking into account the demands of customers. Our furter efforts are connected with the developing a simultaneous bilingual search and extraction features.

The procedure of the role functions extraction centered at the analysis of knowledge structures is sufficiently promising from the point of view of the knowledge bases technology development. The current task is to improve its performance.

The methods were tested on the basis of the summaries of incidents which contain about three thousand documents (each document consists of 10 - 80 lines). In the case of the summaries processing the documents with the mentioned enumerations (there were about 10% of them) were withdrawn and in the remained texts the gaps in the words were removed. At the current moment the program which realizes the proposed procedure gave about 80% of correct recognition of role functions, and about 65% of complete explanations with the indication of all acts. But these numbers rapidly change for the better due to the means (the LK and OFK knowledge) of tuning the LP to special features of the subject area texts. For this not much time is required. Let us note that tuning itself to the extraction of the role functions of persons from the mentioned summaries (with reaching the indicated percentages), required about two weeks of the work of one person. The development and fixing of the shell itself took about four days. The subsequent development is connected with the improvement and the tuning of LP to the work with complex NL forms. At present the extraction of actions is interfered with causal word combinations of the type "*out of the hooligan motives*", "*owing to the hostile relations*" and so forth, which at present are introduced into the system. Difficulties appear with the transfer of the subject of action to other actions to which the subject is not assigned explicitly, but its presence is implied.

The second direction of research and development is connected with the extension of the shell features to the solution of other problems connected with the estimation of objects depending on the nature of statements about them in the texts of description. Within the framework of the studies conducted it is also intended to tune the shell to the work with the English language texts. Since the meaningful portraits

of the English language and Russian language texts have the identical structure (SS-documents), this tuning cannot be labor-consuming.

Bibliography List

[Kuznetsov, 1986] Kuznetsov, I.P. Semanticheskie Predstavleniia. Moscow: Nauka, 1986, 290 p.

[Kuznetsov et al., 2009] Kuznetsov I.P., Efimov D.A., Kozerenko E.B. Tools for Tuning the Semantix Processor to Application Areas // Proceedings of ICAI'09, Vol. I. WORLDCOMP'09, July 13-16, 2009, Las Vegas, Nevada, USA. - CRSEA Press, USA, 2009. P. 467-472.

[Kuznetsov et al., 2009] Kuznetsov I.P., Kozerenko E.B., Kuznetsov K.I., Timonina N.O. Intelligent System for Entities Extraction (ISEE) from Natural Language Texts // Proceedings of the International Workshop on Conceptual Structures for Extracting Natural Language Semantics - Sense'09, Uta Priss, Galia Angelova (Eds.), at the 17 International Conference on Conceptual Structures (ICCS'09), University Higher School of Economics, Moscow, Russia, 2009. P. 17-25.

[Kuznetsov, Kozerenko, 2003] Kuznetsov, I.P., Kozerenko T.B. The system for extracting semantic information from natural language texts // Proceeding of International Conference on Machine Learning. MLMTA-03, Las Vegas US, 23-26 June 2003 г., p. 75-80.

[Banko et al., 2007] Banko M., M. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni. Open Information Extraction from the Web // Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI-07), 2007. P. 2670–2676.

[Clark et al., 2007] Clark P., P. Harrison, and J. Thompson. A Knowledge-Driven Approach to Text Meaning Processing // Proceedings of the HLT-NAACL 2003 Workshop on Text Meaning, 2007. P. 1–6.

[Gildea, Palmer, 2002] Gildea D. and M. Palmer. The necessity of syntactic parsing for predicate argument recognition. In Proceedings of the 40th Annual Conference of the Association for Computational Linguistics (ACL-02), Philadelphia, PA, 2002. P. 239–246.

[Pasca, Van Durme, 2007] Pa¸sca M. and B. Van Durme. What You Seek is What You Get: Extraction of Class Attributes from Query Logs // Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI-07), 2007. P. 2832–2837.

[Punyakanok et al., 2008] Punyakanok V., D. Roth, and W. tau Yih. The Importance of Syntactic Parsing and Inference in Semantic Role Labeling // Computational Linguistics 34(2), 2008. P. 257–287.

[Cunningham, 2005] Cunningham H. Automatic Information Extraction // Encyclopedia of Language and Linguistics, 2cnd ed. Elsevier, 2005.

[Han, 2004] Han J., Pei Y. Yin, and Mao R. Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach," // Data Mining and Knowledge Discovery, 8(1), 2004. P. 53–87.

[Han, Kamber, 2006] Han J. and Kamber, M. Data Mining: Concepts and Techniques // Morgan

ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ ИЗВЛЕЧЕНИЯ СУЩНОСТЕЙ, ОСНОВАННЫЕ НА РАСШИРЕННЫХ СЕМАНТИЧЕСКИХ СЕТЯХ

Кузнецов И.П., Шарнин М.М., Козеренко Е.Б., Мацкевич А.Г., Николаев В.Г., Сомин Н.В.

Учреждение Российской академии наук
Институт проблем информатики РАН,
Москва, Российская Федерация

igor-kuz@mtu-net.ru

В данной работе приведены модель, механизмы и реализации интеллектуальных систем извлечения знаний - сущностей и их связей - из текстов на естественном языке. В основе проектных решений лежит аппарат расширенных семантических сетей. Инструментом создания систем служит оригинальных язык логического программирования – ДЕКЛ.

ВВЕДЕНИЕ

Дается описание семантико-ориентированного лингвистического процессора, извлекающего из текстов естественного языка информационные объекты, их свойства и связи и формирующего на этой основе структуры знаний. Одно из направлений развития таких процессоров связано с выявлением имплицитной информации, которая рассматривается в узком плане - как выявление новых свойств объектов, заданных в неявном виде. Предлагается методика такого выявления, основанная на анализе структур знаний. В качестве примера рассматривается выявление ролевых функций фигурантов на базе их описаний в сводках происшествий.

ОСНОВНАЯ ЧАСТЬ

Научное направление, связанное с обработкой произвольных текстов ЕЯ (в заданной предметной области) развивается с учетом задач определенной категории пользователей: большинство пользователей – специалистов в своей области - интересуются лишь конкретными вещами для своих задач. Например, следователям важны фигуранты, их места жительства, телефоны, приметы, действия лиц (с указанием места, времени), связи и др. Такая информация является основой для оперативно-розыскных действий, поиска по связям, различных видов анализа. Специалистов по кадрам интересуют организации, где человек работал, кем и когда это было. Подобную информацию будем называть *сущностями, информационными объектами*, которые различаются по типам. Например, лица и фигуранты – это объекты одного типа, адреса – другого и т.п.

ЗАКЛЮЧЕНИЕ

Анализ методики проводился с использованием сводки происшествий, содержащей около трех тысяч документов (каждый документ содержит от 10 до 80 строк). Предлагаемая методика, основанная на анализе структур знаний, представляется перспективной для развития технологии баз знаний.