Linguistic Processor “Semantix” for Knowledge Extraction from Natural Texts in Russian and English

Igor P. Kuznetsov
Institute for Informatics Problems of the Russian Academy of Sciences
Moscow, Russia
igor-kuz@mtu-net.ru

Elena B. Kozerenko
Institute for Informatics Problems of the Russian Academy of Sciences
Moscow, Russia
elenakozerenko@yahoo.com

Abstract The linguistic processor Semantix is intended for the areas where the automatic formalization of the flows of texts in natural language is required: resume, mass media issues, information and advertising materials, mail communications, summaries of incidents, information in the criminal cases, archive materials and other texts. The objects interesting for a user are extracted from documents with their features and relations. As a result on the basis of each document a special form of the semantic network is built, which reflects its semantic structure. Such networks are mapped onto the XML-files, which serve for organizing the bases of knowledge, corresponding to semantic search, for the solution of logical analytical problems, and also for the automatic filling of relational databases

Keywords: semantics, natural language, linguistic processor, knowledge engineering, data extraction

1 Introduction

The studies of the chief analysts show that the joint volume of digital information in 2006 was 161 million gigabytes. It is assumed that during the period from 2006 to 2010 the volume of information will increase in more than six times. In more than 80% of cases this information is unstructured - these are the texts of natural language. For humans the difficulties with orientation in the flows of incoming information increase dramatically. In connection with information processing the new innovation approaches, oriented at the tasks of concrete users, are required. One should consider that the large category of users have the specific official responsibilities, and respectively, constant interests. Completely concrete information is necessary for them. For example, a criminal inspector seeks to extract information on important figurants, their places of residence, telephones, criminal events, dates and other such facts; a personnel manager is interested in the organizations, when and where a person worked and in what position. Other people try to fish out from the media the information about the countries, important persons, catastrophes, etc. We call this concrete information interesting for a user information objects. Objects are distinguished by their types.

Hence follows the need for constructing a new class of information systems, which would consider the interests of the end user and be oriented at extracting information objects from texts [1,11-17]. At present this problem is in the focus of attention of many researchers and developers [2-19].

In this article a class of such systems is presented, based on the use of special linguistic processors (LP) and technology of knowledge bases (KB). Linguistic processors are necessary for the deep processing of texts with the development of information objects and connections. On the basis of the latter the structures of the knowledge comprised in the knowledge base are formed. We call such processors semantics-oriented. Their special feature is the employment of the linguistic knowledge (LK), organized in such a way as to consider lexical and semantic special features of natural language with the formation of the knowledge structures [1,14]. At the level of KB it is possible to consider more fully the needs of the users for decision of the following tasks.

First, due to the use of the reverse linguistic processors the formation of reports, filling the required table forms and relational databases have become possible.

Second, due to the support of the expert component, it is possible to ensure the updating of the information by the analytical results, obtained via processing the knowledge structures.

Third, intelligent features are provided due to the organization of different types of search: the search for concrete objects, the search for similar objects, the search for connections, etc. Such forms of search relate to the "semantic" facilities, since the results are achieved not at the level of words or word forms, but at the level of the knowledge structures from KB. We call the systems of this type semantics-oriented.

During the last fifteen years on the basis of the studies conducted at the Institute for Informatics Problems of the Russian Academy of Sciences the semantics-oriented systems and linguistic processors have been developed for the formalization of natural language texts and their analytical processing for different subject areas: criminology (summary of incidents, accusatory conclusions, etc.), the Media (documents about terrorist activities), personnel
management (autobiographies in the Russian and English languages). These are integrated systems DIEZ, IKS, "Analyst, "Criminal", Lingua-Master [12-17]. Their scientific base is comprised by the extended semantic networks (ESN), the methods of presentation of complex knowledge configurations, the instrumental environment (DEKL) for processing the structures of knowledge, network position grammars, ontologies in the form of ESN, morphological analysis on the basis of the generalized inflectional endings [1,12-14,19]. The last version of this processor in the form of the SDK module, developed in collaboration with the Synergistic Systems software company is called Semantix.

This paper presents a discussion of special features of this system, the linguistic processors and knowledge bases employed in them determined by the tasks and specific character of natural language.

2 Basic components of Semantix

The linguistic processor Semantix is intended for the areas, where the automatic processing of the flows of natural language (NL) texts is required. On the basis of each document a special form of the semantic network (ESN) is built, which presents its semantic structure. This network is mapped onto XML-file. With their aid the subsequent automatic analysis is considerably facilitated. XML files are a basis for the composition of dossier, surveys, reports. Different versions of these systems use either an automatic filling relational database (DB) or the forming of their own knowledge base with the subsequent organization of directed retrieval for the necessary information (objects), including different forms of semantic search.

1.1. The block of lexical and morphological analysis. It extracts words and sentences from the text, gives the words their normal form and constructs the semantic network, presenting the space structure of text (SST), which reflects the sequence of words, their basic features, beginnings of sentences and presence of space characters lines. The block uses a special collection of subject dictionaries (the dictionary of countries, regions of Russia, names, forms of weapon and others) for grouping the words and giving them additional semantic attributes [19].

1.2. The block of syntactic-semantic analysis. It converts one semantic network (SN) into another, represented semantic structure of text (SST) the, i.e., the chosen objects and their connections. The latter is frequently called the meaningful portrait of document [11,12]. The block is governed by the linguistic knowledge (LK), which determine the process of text analysis. LK include special form of contextual rules, which ensure the high degree of selectivity with the development (extraction) of objects and connections [14]. The tasks of this block are as follows:

- Extraction of the information objects from the flow of NL-documents: persons, organizations, actions, their place and time, and many other objects.
- The establishment of the connections of objects. For example, persons are connected with the organizations (PLACE_OF_WORK), by the addresses (LIVES, REGISTERED). Or figurants of criminal cases are connected with such objects as the type of weapon, drugs (TO HAVE).
- The analysis of finite and nonfinite verbal forms with the identification of the participation of objects in the appropriate actions. For example, one figurant gave drugs to another figurant, and this is the fact, binding them - The establishment of the connections of actions with the objects by place or time (where and when some action or event occurred).
- The analysis of the reason-consequence and temporary connections between actions and events.

1.3. Expert systems (ES). On the basis of semantic networks the new knowledge pieces are constructed in the form of additional fragments (ESN). For example, the ES extracts the area of the person activity (in accordance with the assigned classifier) from the text of resume for each autobiography. The experience of the person’s work is evaluated. The correlation of a criminal incident to the specific type is accomplished with the analysis of the criminal actions of ES: the following facts are revealed - the nature of crime, the method of its accomplishment, the instrument, and so forth (in accordance with the classifiers of the criminal police).

1.4. Reverse linguistic processor, which converts the meaningful portrait of document (semantic network) into the XML-file. In this case the necessary replacements of symbols, service words (names of objects) are achieved, the markers of beginning and end of the objects, actions, sentences. Conversion is achieved without the loss of information. The XML-file is arranged in such a way that all the revealed components and connections are represented in it. If necessary, the inverse transformation of the XML-file into the semantic network is ensured.

1.5. The base of linguistic and expert knowledge (KB). It contains the rules of the text analysis and expert solutions in the internal presentation. They determine the work of the linguistic processor. Semantix has several such bases, which are activated depending on subject areas and user tasks.

2. The objects and links for extraction

The set of the objects to be extracted depends on the tasks of a user. At the same time the quality of a linguistic processor is to a considerable degree determined by the possibilities for this extraction. The basic types of information objects and connections, extracted by Semantix [21] are enumerated below:

- persons (by family name, given name and patronymic - FNP) with their role features (criminal, victim);
- the verbal description of the persons, their distinctive signs;
- address, posting information attributes;
- date(s) mentioned;
- weapon with its special features;
- telephone numbers, faxes, e-mails with their subsequent standardization;
- the means of transport with the indication of the vehicle type, its state number, color and other attributes;
• passport data and other documents with their attributes;
• explosives and narcotic substances;
• organizations, positions;
• quantitative characteristics (how many persons or other objects participated in an event);
• the numbers of accounts, sums of money with the indication of the currency type;
• terrorist groups and organizations;
• participants of terrorist groups with the indication of their roles (leader, head of, etc.);
• the armed forces, assigned for antiterrorist combat (MilitaryForce);
• event (criminal, terrorist, biographical, and so on) with the indication of the information objects participation in them;
• time and the place of events;
• the connection between different types of information objects (with whom a person works in an organization, or lives at the same address, in what events participated together with other objects, etc.).

Figure 1 represents an example of the set of selected objects in the demonstration version of the Semantix processor [21].

For extracting objects all versions of an object name including the brief form possible in the text were considered. Standard objects (names, dates, addresses, the forms of weapon and others) are reduced to one (standard) form. The identification of objects is performed taking into account brief designations (for example, separate surnames, patronymics, initials), anaphoric references (indicative and personal pronouns, for example, "this person", "it...") definitions and explanations (for example, "the mayor of Moscow Luzhkov" is identified with the subsequent words "mayor", "Luzhkov"). For the extraction of events and connections the analysis of verbal forms, participial and adverbiai constructions is carried out.

An important task is the identification of objects in the entire text, the use for these purposes of indicative pronouns, brief names, anaphoric references. Taking into account the difficulties and in accordance with the tasks the linguistic processor Semantix was developed, which achieves normalization of words, their grouping with the formation of units, the identification of objects and the establishment of connections. As a result for each NL document a semantic network called the meaningful document portrait was constructed automatically. The latter are the knowledge structures of the knowledge base which serve the basis for implementing different forms of semantic search: the search by features and connections, the search for the objects connected at different levels, the search for similar figurants and incidents, the search by distinctive signs (with the use of ontologies).

The extraction of connections is not only the deep analysis of verbal and other forms. Many connections are given on default. For example, in the summaries of incidents, as a rule, figurants names are followed by their data without the indication of their belonging and with the additional text insertions. For that the directed search for the connected objects, i.e., the restoration of connections, default data is organized in the processor Semantix. Special processes are organized in order to connect persons with their place of stay or place of work, vehicles which belong to them, and so forth. For example, the analysis of the summaries of incidents is performed as follows. For a number of objects (address, telephone, date of birth, etc.) a virtual connection with other objects (names, organizations), is built thus yet unidentified. Then, at the same level of processing their search is performed with the aid of the special rules for identification. In these rules the direction of search, the permissible quantity of steps, and also the signs of words and punctuation marks, where the process of search ends are indicated. In this case special filters are required, in order not to take and not to connect an alien object. This approach showed sufficiently good results in the system Criminal [16]. The special features of natural language are considered where the same actions are identified with the aid of the verbs, verbal nouns and participial constructions. Presented in ESN they are reduced to one form, i.e. a complex object. Moreover, forms with verbal nouns can be the components of verbal forms. On analogy, in ESN some objects can be the components of others. The reason-consequence and temporary dependences between actions, events, etc. are represented which reflect the logical connection of sentences, assigned explicitly, with the aid of the words “therefore”, “then”, etc. Another example of meaningful portrait, presented in the form of a graph, is represented in Figure 2.

3 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 language 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.

Subject areas
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 [19]. Note that in the mentioned processors the objects are brought to the standard form (for example, FNP, address, date) with the indication of the types of components. 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 [21]. At present the first release of the English language version of the object - oriented linguistic processor Semantix [15,16,19, 21] has been developed.

5 The structure of the XML file

In the XML file a meaningful portrait of a document (the semantic network structure) is represented comprising all objects and connections, revealed by the Semantix text processor. In connection with this the organization of XML files has the definite scientific value as the means for presentation of the semantic structure of sentences and texts. The transformation of the semantic network into the XML file is ensured with the aid of the reverse linguistic processor. In this case the fragments which present objects, relations, actions and sentences in the semantic network structure are mapped onto the appropriate components of the XML file which will also contain objects, relations, actions and sentences.

The basic task of the LP use consists in operation as a separate module within the framework of the integrated systems of information collection and processing. The exchange is conducted through XML files [20]. For that end a reverse LP was developed, which constructs XML files on the basis of meaningful portraits.

Thus, the input for the linguistic processor (LP) is a natural language text, and the output is an XML- file, where all chosen objects and connections with the indication of sources are represented. This LP named Semantix is provided in the form of an SDK- module. It works under WINDOWS, but it can be recompiled for the work under LINUX.

The Semantix Processor is an independent module and it can be used without the mentioned systems for the standard tasks of analytical services. There are means of tuning to the objects of other types - due to the linguistic knowledge or the dictionaries.

Let us give some explanations. Each object has the following structure:

```xml
- <OBJECT ID="5" TYPE="WEAPON">
  <ARG TEXT="TROTYL" />
  <ARG TEXT="BLOCK" />
  <ARG TEXT="OF" />
  <ARG TEXT="200" />
  <ARG TEXT="GRAM" />
  <SOURCE>Trotyl block of 200 grams</SOURCE>
</OBJECT>
- <OBJECT ID="6" TYPE="PLACE">
  <ARG TEXT="SETTLEMENT" />
  <ARG TEXT="KHVANI" />
  <ARG TEXT="" />
  <ARG TEXT="KUNDA" />
  <SOURCE>Settlement Khvani Kunda</SOURCE>
</OBJECT>
- <OBJECT ID="7" TYPE="TERRORIZM">
  <ARG TEXT="CHECHEN" />
  <ARG TEXT="BAND" />
  <SOURCE>Chechen bands</SOURCE>
</OBJECT>
- <OBJECT ID="8" TYPE="FIGO">
  <ARG TEXT="KHATTAB" TYPE="SURNAME" />
  <ARG TEXT="HASAN" TYPE="NAME" />
  <SOURCE>Hasan Khattab</SOURCE>
</OBJECT>
```

Figure 3. An example of XML file for the semantic structure presentation.

[9] 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, pp. 53–87.