Intelligent System for Entities Extractions (ISEE) from Natural Language Texts

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
kozerenko@mail.ru

Konstantin I. Kuznetsov
Synergetics Systems
Moscow, Russia
k.kuznetsov@synsys.ru

Natalia O.Timonina
Institute for Informatics Problems of the Russian Academy of Sciences
Moscow, Russia
lavespa@mail.ru

Abstract

This paper describes a semantic linguistic processor which extracts the entities and their links from natural language texts. The conceptual model underlying the algorithmic developments is the extended semantic networks (ESN). This paper analyzes the use of the processor for text formalization in various subject fields: economy monitoring, criminal actions, mass media, terrorist activities (in Russian and English). Peculiarities of the texts are taken into account by linguistic knowledge of the processor: the system can be tuned to various subject areas. We describe the use of this processor for text formalization in different subject areas, such as economic crisis monitoring, criminology (summary of incidents, accusatory conclusions, etc.), the mass media documents about terrorist activities, personnel management (autobiographies, resume). Special features of each problem area are examined: the collections of extracted entities, the means for their identification, their connections, occurring contractions, punctuation and special signs, specific character of language constructions, etc. - all these special features were taken into account in the linguistic knowledge development.

1 Introduction

A tremendous increase of document resources, obtained by the users through different information channels (including the Internet), requires new solutions. The analytical reports predict that during the period from 2006 to 2010 the volume of information will increase in more than six times. The majority of such documents (about 80%) exist in the form of natural language (NL) texts. It is impossible to read and comprehend even the smallest portion of the factual information available. The existing information systems can render assistance, but for this a preliminary formalization is required. At the same time a great number of end users are people interested in specific subject. For example, a financial analyst is interested in facts, figures and references relating to banks, stock and shares, economic situation and dynamics of different companies and markets; a criminal inspector seeks to extract information on important suspects of criminal acts, 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 extract from the media the information about countries, cities, places of interest, monuments, important persons, catastrophes, etc. We call these particular objects of user interest *entities*. Entities together with their features and links form *information objects.*

Hence follows the need for constructing a new type of information systems, which would consider the interests of the end user and be oriented at extracting information objects from texts [1-3]. At present this problem is in the focus of attention of many researchers and developers [4-19].

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 *entities* 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 deciding 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 of knowledge structures.

Third, intelligent features are provided due to the organization of different types of search: the search for concrete entities, the search for similar entities, 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 Intelligent Systems for Entities Extraction (ISEE). ISEE are *semantics-oriented.*

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

2 The Extended Semantic Networks Presentation Mechanism

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). The linguistic processors under consideration were implemented within the framework of a series of the knowledge extraction and processing systems DIEZ, IKS, "Analyst, "Criminal", Lingua-Master, LOGOS_D [12-17] at the Institute for Informatics Problems of the Russian Academy of Sciences. These systems are integrated intelligent environments with natural language interfaces, designed for knowledge extraction from unprepared ("raw") texts and knowledge management. On their basis a number of application systems for specific problem domains were developed. Though the majority of applications were made in Russian, the multilingual features have been developed from the very beginning. The implementation platform was the DECLAR logical programming language (and programming environment) designed in the Microsoft paradigm. The DECLAR language is a special development featuring the uniform mechanisms for knowledge structures presentation and processing. The logic of the DECLAR language resembles the logic of PROLOG, but it is a different tool which enables a system designer to employ several inference mechanisms and supports the predicate structures with an arbitrary number of arguments.

At present a new domain is being introduced, that is economic crisis monitoring. Originating from the projects with deep semantic approach to linguistic knowledge presentation, the models described here are interlingua-based. In our case the interlingua is a language of extended semantic networks which are predicate expressions supporting the mechanism of embedded structures. The simulation objective was to work out implementable semantic invariants for syntactic structures in a subset of natural languages (English and Russian in this case) for the purposes of cross-lingual electronic communications connected with the process of knowledge exchange, management and a focused document translation. The predicate expressions in our systems form a sentence frame and can take sentential arguments. Constructions with nonfinite verb forms are considered as the result of the syntactic derivation process which is very similar in the examined languages. The resulting presentation for each verbal vocabulary item is supplied with a syntactic derivation history which is given as a juxtaposition of primary, secondary, etc. meanings. The resulting shifts of semantic-syntactic case frame structures with all possible types of arguments are presented as well. This presentation serves as a source of constraints for syntactic analysis algorithms in multilingual processors of the knowledge management systems based on extended semantic networks (ESN).

All knowledge pieces in the system including the linguistic knowledge are semantic network fragments, i.e. named predicate structures, and the knowledge processing procedures are production rules which modify and transform the semantic network. Thus the linguistic processor in our case is a declarative knowledge base section. Besides the processor, the dictionaries for each included language are supported, and the invariant semantic dictionary containing the "sense" codes without any language-specific information.

Language-specific dictionaries contain lexical and morphological descriptions of words, the syntactic structures types and the rules are contained in the linguistic processor. The design solutions envisage the facility of language extensions.

A special linguistic processor Semantix has been designed within the DECLAR environment. Semantix is an intelligent tool for knowledge discovery in the flows of natural language texts. A special form of the *extended semantic network* (ESN) is built on the basis of each text document. The semantic network presents the document semantic structure including entities, relations, links and implications. This network is automatically mapped into an XML file. XML files serve as the basis for compilation of analytical surveys, reports, dossiers, etc. The intelligent tool is multifunctional and can be reconfigured for particular applications. Different versions of intelligent Semantix environments have been implemented. These systems can be used for automatic population of relational databases (DB) from natural language texts, for creation of knowledge bases with subsequent organization of directed retrieval to the necessary information (semantic entities).

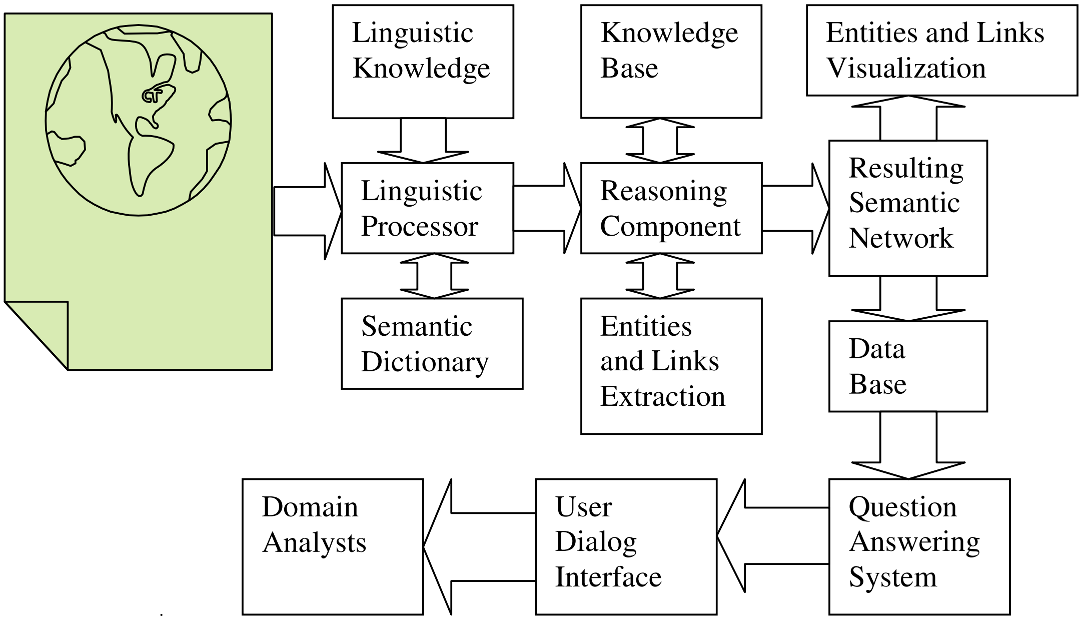


Figure 1. General Functions of the Intelligent System for Entities Extractions (ISEE).

3 Semantix Components

A variety of semantic search forms is supported. The linguistic processor Semantix comprises the following components.

1). *The component of lexical and morphological analysis*. It extracts words and sentences from the text, performs lemmatization of words (normal form establishment) and constructs the semantic network presenting the space structure of text (SpST), which reflects the sequence of words, their basic features, beginnings of sentences and the presence of space character lines. The component uses a two-level general ontology and a special collection of subject dictionaries (the dictionary of countries, regions of Russia, names, forms of weapons, and other items specific for the supported domains). The component performs semantic grouping of the words and assigns them additional semantic attributes.

2). The *component of syntactic-semantic analysis*. It converts one semantic network (SN) into another one which represents the semantic structure of the text (SemST), i.e., the relevant semantic entities and their connections. The SemST is called *the meaningful portrait of document*. It comprises knowledge structures of the knowledge base which serve the as basis for implementing different forms of semantic search : the search by features and connections, the search for the entities connected at different levels, the search for similar persons and incidents, the search by distinctive characteristics (with the use of ontologies).

The component is controlled by the linguistic knowledge (LK), which determines the process of text analysis. LK includes special form of contextual rules which ensure a high degree of selectivity with the extraction of entities and connections. The functions of this component are as follows:

- Extraction of information objects from the flow of NL documents: persons, organizations, actions, their place and time, and many other relevant types of objects.
- The establishment of connections between objects. For example, persons are connected with organizations (PLACE_OF_WORK), by addresses (LIVES, REGISTERED). Or criminals and suspects 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 criminal gave drugs to another criminal, and this is the fact linking 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.

3). *Expert system component (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 a 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).

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.

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.

4 The Entities and Links for Extraction

The set of the entities 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 Semantix processor supports more than 40 types of semantic entities which can be extracted automatically. Some examples of basic entities types and connections extracted by Semantix are given below:

- persons (by family name, given name and patronymic[1] - 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 license number, color and other attributes;

[1] In Russian it is a middle name formed from one's father's name observing special morphology, e.g. Ivan – Ivan*ovich* (for men), Ivan*ovna* (for women); patronymics are a regular way of addressing politely adult people, e.g. Ivan Ivanovich, or Maria Ivanovna; thus, the full name of a Russian person consists of a given name, a patronymic and a family name, e.g. Ivan Ivanovich Ivanov.

- 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 (Military_.Force);
- 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.).

For extracting objects all versions of an object name including the contracted form possible in the text were considered. Standard objects (names, dates, addresses, types of weapons 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 adverbial constructions is carried out.

An important task is the identification of entities in the entire text, the use for these purposes of indicative pronouns, brief names, anaphoric references.

5 Factors of Processor Quality

The quality of a linguistic processor is determined by a number of factors.

The first one is the facility of entities and connections establishment. The Semantix processor outperforms the existing systems by the number of the supported semantic entities types. It identifies more than 40 types of entities including very complex ones, which correspond to actions and events, for comparison the competitors’ best result is about 15 types.

The second important factor is the selectivity of rules and procedures of identification: the factor of the noise and losses. By noise we mean the presence of excessive words in the entities. Losses are the situations when an entity is not revealed or revealed partially: in the text there are the words, which did not enter into the entity. 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 entities 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 entities), and also tuning to the new subject domains and types of entities. Due to the complexity of analysis this tuning is achieved through the linguistic knowledge (LK). 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 entities being extracted. The application of rules of extraction is connected with the search for the necessary words, where sorting is required. The greater the number of entities and rules the greater is the time of analysis. In the Semantix processor there are different means of decreasing sorting time. Besides the program, there are also means of control by linguistic knowledge. It is indicated for each rule, what words should be included for the initiation of the process of its application. The constraints in the form of the expected contexts (to the left and to the right of the 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 entities extracted.

The system features complex means which ensure rapid tuning to the applications (including the introduction of new entities and connections) taking into account the demands of customers. Note that in the mentioned processors the entities are brought to the standard form 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 entities of the entire text. The analysis of 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 information object - oriented linguistic processor Semantix has been developed.

6 The On-Going Developments

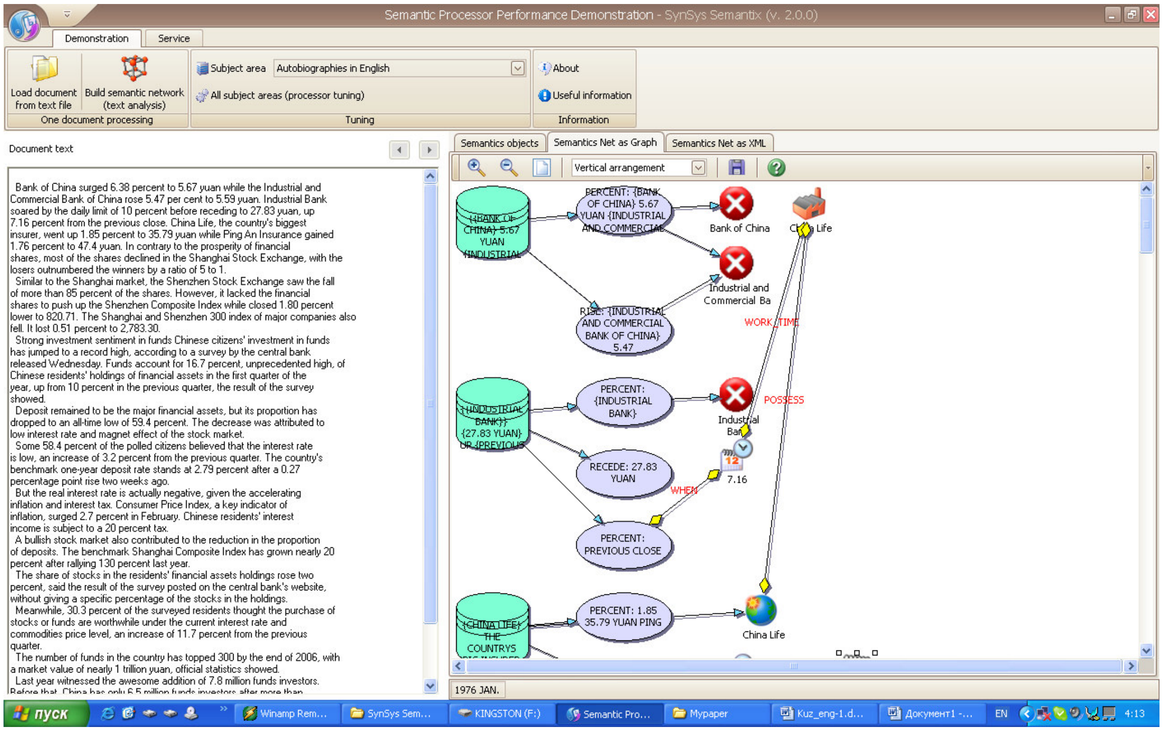
6.1 Economic Monitoring

The English release of the Semantix linguistic processor is being tuned to the Financial knowledge extraction and analysis and semantic Web search for the relevant information in Economics sphere. This release is intended for intelligent support of business solutions. The primary investigation stage requires cognitive linguistic analysis of the subject area with the subsequent design of the domain-oriented ontology and its incorporation into the general ontology of Semantix. The development of the domain-specific vocabularies can be partially supported by the learning facilities of the linguistic processor.

The system performs semantic objects extraction from natural language texts and automatic formation of semantic networks. The networks are populated by objects of different types and their interconnections: banks, companies, addresses, stock, shares, values, stock exchanges, etc. The objects are extracted from documents with their features and relations. On the basis of each document a special form of semantic network is built reflecting its semantic structure. Such networks are mapped into XML- files which constitute knowledge bases. Semantic search results in semantic presentations for deciding logical analytical problems and for automatic filling of relational databases.

Three modes of semantic presentations are supported by the system:

1. List of objects types with references to particular content items, e.g. Type: Mentioned Organizations; Content: Bank of China; Source Text: "Bank of China";
2. Graph presentation of semantic network constructed from the texts (Fig.2);
3. XML presentation of semantic network constructed from the texts.


Semantic Processor Performance Demonstration - SynSys Semantix (v. 2.0.0)
Demonstration
Service
Load document from text file
Build semantic network (text analysis)
One document processing
Subject area
Autobiographies in English
All subject areas (processor tuning)
Tuning
About
Useful information
Information
Document text
Semantics objects
Semantics Net as Graph
Semantics Net as XML
Vertical arrangement
{{BANK OF CHINA} 5.67 YUAN {INDUSTRIAL
PERCENT: {BANK OF CHINA} 5.67 YUAN {INDUSTRIAL AND COMMERCIAL
Bank of China
China Life
Industrial and Commercial Ba
RISE: {INDUSTRIAL AND COMMERCIAL BANK OF CHINA} 5.47
WORK_TIME
{{INDUSTRIAL BANK}} {27.83 YUAN} UP {PREVIOUS
PERCENT: {INDUSTRIAL BANK}
Industrial Ba
POSSESS
RECEDE: 27.83 YUAN
7.16
WHEN
PERCENT: PREVIOUS CLOSE
{CHINA LIFE} THE COUNTRYS
PERCENT: 1.85 35.79 YUAN PING
China Life
1976 JAN.
пуск
Winamp Rem...
SynSys Sem...
KINGSTON (F:)
Semantic Pro...
Mypaper
Kuz_eng-1.d...
Документ1 -...
EN
4:13

Bank of China surged 6.38 percent to 5.67 yuan while the Industrial and Commercial Bank of China rose 5.47 per cent to 5.59 yuan. Industrial Bank soared by the daily limit of 10 percent before receding to 27.83 yuan, up 7.16 percent from the previous close. China Life, the country's biggest insurer, went up 1.85 percent to 35.79 yuan while Ping An Insurance gained 1.76 percent to 47.4 yuan. In contrary to the prosperity of financial shares, most of the shares declined in the Shanghai Stock Exchange, with the losers outnumbered the winners by a ratio of 5 to 1.

Similar to the Shanghai market, the Shenzhen Stock Exchange saw the fall of more than 85 percent of the shares. However, it lacked the financial shares to push up the Shenzhen Composite Index while closed 1.80 percent lower to 820.71. The Shanghai and Shenzhen 300 index of major companies also fell. It lost 0.51 percent to 2,783.30.

Strong investment sentiment in funds Chinese citizens' investment in funds has jumped to a record high, according to a survey by the central bank released Wednesday. Funds account for 16.7 percent, unprecedented high, of Chinese residents' holdings of financial assets in the first quarter of the year, up from 10 percent in the previous quarter, the result of the survey showed.

Deposit remained to be the major financial assets, but its proportion has dropped to an all-time low of 59.4 percent. The decrease was attributed to low interest rate and magnet effect of the stock market.

Some 58.4 percent of the polled citizens believed that the interest rate is low, an increase of 3.2 percent from the previous quarter. The country's benchmark one-year deposit rate stands at 2.79 percent after a 0.27 percentage point rise two weeks ago.

But the real interest rate is actually negative, given the accelerating inflation and interest tax. Consumer Price Index, a key indicator of inflation, surged 2.7 percent in February. Chinese residents' interest income is subject to a 20 percent tax.

A bullish stock market also contributed to the reduction in the proportion of deposits. The benchmark Shanghai Composite Index has grown nearly 20 percent after rallying 130 percent last year.

The share of stocks in the residents' financial assets holdings rose two percent, said the result of the survey posted on the central bank's website, without giving a specific percentage of the stocks in the holdings.

Meanwhile, 30.3 percent of the surveyed residents thought the purchase of stocks or funds are worthwhile under the current interest rate and commodities price level, an increase of 11.7 percent from the previous quarter.

The number of funds in the country has topped 300 by the end of 2006, with a market value of nearly 1 trillion yuan, official statistics showed.

Last year witnessed the awesome addition of 7.8 million funds investors. Before that, China has only 6.5 million funds investors after more than

Fig. 2. Graph presentation of semantic network constructed from the "Bank of China" text

6.2 Monuments Catalogs Intelligent Guide

The Intelligent Guide for extraction of particular information about monuments and historical places of interest in under development at present. The principal entities extracted by this system are historical monuments, buildings, persons and events connected with these monuments; historical periods and dates. The system supports the feature of identification of different monuments with the particular person to whom these monuments were dedicated, as, for example, the following descriptions of different monuments will be linked with the same historical person – tsar Peter I:

the bronze copy of the original "Tsar Carpenter" monument remained in Holland.

"The Bronze Horseman" is the most famous monument of St. Petersburg.

The monument was built between 1845 and 1847. On May 23, 1848 it was opened.

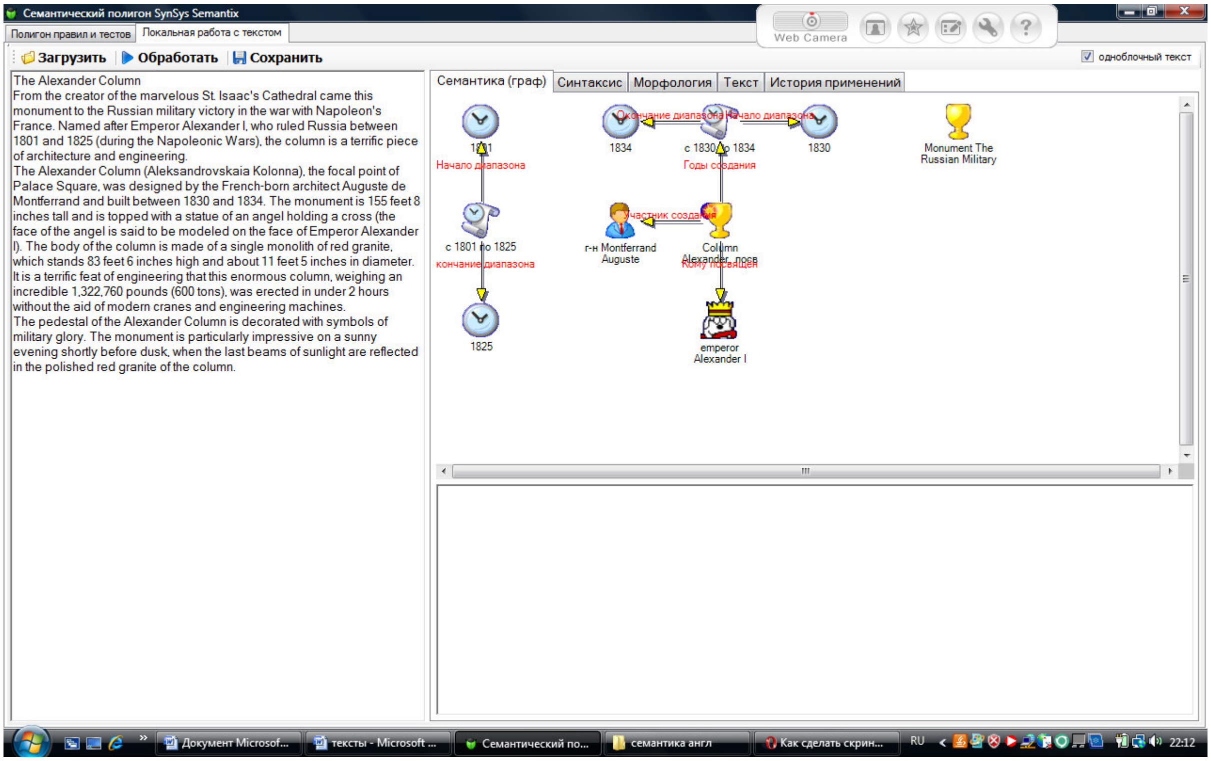
In 2000 the statue was installed and in 2001 it was destroyed.

Figure 3 shows an example of the Intelligent System for Entities Extraction (ISEE) populated from the text about the Alexander Column in St. Petersburg.

6.3 Related Work and Evaluation

The developments described in this paper relate to the field of Artificial Intelligence and, hence share the aspirations of related research and development projects which are aimed at factual knowledge presentations and natural language processing. The developments which in certain aspects are closest to ours and have similar features are FASTUS and Cyc [4,22-24]. The Cyc project (started in 1984 by Douglas Lenat and developed by company Cycorp) is an artificial intelligence tool that assembles a comprehensive ontology and knowledge base of everyday common sense knowledge, with the goal of enabling AI applications to perform human-like reasoning. The presentation mechanisms of Cyc resemble Lisp structures, natural language processing is oriented at the English language.

At present we develop the uniform methods of evaluation and performance comparison for the AI systems featuring the named entity extraction and natural language analysis.



The Alexander Column
From the creator of the marvelous St. Isaac's Cathedral came this monument to the Russian military victory in the war with Napoleon's France. Named after Emperor Alexander I, who ruled Russia between 1801 and 1825 (during the Napoleonic Wars), the column is a terrific piece of architecture and engineering.
The Alexander Column (Aleksandrovskaia Kolonna), the focal point of Palace Square, was designed by the French-born architect Auguste de Montferrand and built between 1830 and 1834. The monument is 155 feet 8 inches tall and is topped with a statue of an angel holding a cross (the face of the angel is said to be modeled on the face of Emperor Alexander I). The body of the column is made of a single monolith of red granite, which stands 83 feet 6 inches high and about 11 feet 5 inches in diameter. It is a terrific feat of engineering that this enormous column, weighing an incredible 1,322,760 pounds (600 tons), was erected in under 2 hours without the aid of modern cranes and engineering machines.
The pedestal of the Alexander Column is decorated with symbols of military glory. The monument is particularly impressive on a sunny evening shortly before dusk, when the last beams of sunlight are reflected in the polished red granite of the column.

Fig. 3. The ISEE populated by the objects from the “Alexander Column” text.

7 Conclusion

The systems based on the entities extraction linguistic processors can be used in different areas of application where the extraction of useful information from natural language texts is required. In this case, the processors, described in this work, possess a number of essential advantages.

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 3%.

For performance increase the representative collections of test documents are extremely important. 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]. It has a number of essential advantages, the main one is that it outperforms recently developed other 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 [19]. At present the English language version of the object - oriented linguistic processor Semantix [15,16,19,21] has been released. Our further efforts are connected with the developing of a simultaneous bilingual search and extraction features.

References

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

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

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

[4] FASTUS:a Cascaded Finite-State Trasducerfor Extracting Information from Natural-Language Text. // AIC, SRI International. Menlo Park. California, 1996.

[5] Ferrucci, D. and Lally, A. UIMA: an architectural approach to unstructured information processing in the corporate research environment // Natural Language Engineering 10 (3/4), 2004, 327–348.

[6] Byrd, R. and Ravin, Y. Identifying and Extracting Relations in Text // 4th International Conference on Applications of Natural Language to Information Systems (NLDB). Klagenfurt, Austria, 1999.

[7] Popov, B. et al. KIM - A Semantic Platform for Information Extraction and Retrieval // Journal of Natural Language Engineering, 10(3-4), 2004, pp. 375-392.

[8] Doddington, G. et al. Automatic Content Extraction (ACE) program - task definitions and performance measures // Fourth International Conference on Language Resources and Evaluation (LREC), 2004.

[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.

[10] Dong, G. and J. Li. Efficient mining of emerging patterns: Discovering trends and differences // Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and DataMining, S. Chaudhui and

D. Madigan, editors, ACM Press, San Diego, CA, 1999, pp. 43–52.

[11] Kozerenko, E.B. Multilingual Processors: a Unified Approach to Semantic and Syntactic Knowledge Presentation. In Proceedings of the International Conference on Artificial Intelligence IC-AI'2001. H.R. Arabnia (ed.), Las Vegas, Nevada, USA, June 25-28, 2001. CSREA Press, 2001, pp.1277-1282.

[12] Kuznetsov I.P. Methods of Processing Reports with the Extraction of Figurants and Events Features // In Dialogue'99: Proceedings of the International Workshop "Computational Linguistics and its Applications", Vol.2, Tarusa, 1999.

[13] Kuznetsov I.P., Matskevich A.G. The System for Extracting Semantic Information from Natural Language Texts // Proceedings of the Dialog International Workshop "Computational Linguistics and its Applications", Vol.2, Moscow: Nauka, 2002.

[14] Kuznetsov I.P. Natural Language Texts Processing Employing the Knowledge Base Technology // Sistemy i Sredstva Informatiki, Vol.13, Moscow: Nauka, 2003, pp. 241-250.

[15] Kuznetsov, I., Kozerenko, E. 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, p. 75-80.

[16] Kuznetsov I.P., Matskevich A.G. The English Language Version of Automatic Extraction of Meaningful Information from Natural Language Texts // Proceedings of the Dialog-2005 International Conference "Computational Linguistics and Intelligent Technologies", Zvenigorod, 2005pp. 303-311.

[17] Kuznetsov I.P., Matskevich A.G. Semantics Oriented Linguistic Processor for Automatic Formalization of Autobiographical Data // Proceedings of the Dialog-2006 International Conference "Computational Linguistics and Intelligent Technologies", Bekasovo, 2006, pp. 317-322.

[18] Voss, S. and Joslyn C.A. Advanced Knowledge Integration in Assessing Terrorist Threats // LANL Technical Report LAUR 02-7867, 2002.

[19] Somin N.V., Solovyova N.S., Charnine M.M The System for Morphological Analysis: the Experience of Employment and Modification // Sistemy i Sredstva Informatiki, Vol. 15 Moscow: Nauka, 2005, pp. 20-30.

[20] Gardner, J. R. and Z. L. Rendon, XSLT and XPATH: A Guide to XML Transformations, Prentice Hall, 2001.

[21] Web site with the demo version of the Semantix system: http://semantix4you.com

[22] Lenat, Douglas. "Hal's Legacy: 2001's Computer as Dream and Reality. From 2001 to 2001: Common Sense and the Mind of HAL". Cycorp, Inc.,: http://www.cyc.com/cyc/technology/halslegacy.html

[23] Cyc R&D: http://www.cyc.com/cyc/cycrandd/areasofrandd_dir/is

[24] Chris Deaton et al. (2005). "The Comprehensive Terrorism Knowledge Base in Cyc". In: Proceedings of the 2005 International Conference on Intelligence Analysis, McLean, Virginia, May 2005.