Knowledge acquisition based on natural language texts

Hubarevich Nastassia
Belarussian State University of Informatics and Radioelectronic
Minsk, Belarus
stasia@tut.by

Boyko Igor
Center for System Analysis and Strategic Studies of the National Academy of Sciences of Belarus
Minsk, Belarus
igor_m_boyko@hotmail.com

Semenyaka Anatoly
Belarussian State University of Informatics and Radioelectronic
Minsk, Belarus
toli44777@gmail.com

Hardzei Aliaksandr
Center for System Analysis and Strategic Studies of the National Academy of Sciences of Belarus
Minsk, Belarus
alieks2001@yahoo.com

***Abstract*—The article describes a method of computer analysis of natural language texts and automatic filling the knowledge base using OSTIS technology. The method helps to implement semantic-syntactic analysis of texts and then analyse its context based on specific subject domain ontologies.**

***Keywords*—natural language processing, universal semantic code, text understading**

A cognitive approach in the field of artificial intelligence is under intensive development and it involves modeling various aspects of text understanding [1].

According to D.Pospelov: an intellectual system understands some text if able to answer questions about its content with an accent on the deep semantics but not just pure facts. The sense of the text reflects knowledge represented by a formal language as a semantic equivalent where objects and relations between them are not limited with linguistic categories and where they represent real world objects and relations [2].

I. Introduction

Main components of natural language text analysis are syntactic and semantic.

Syntactic analysis describes a syntactic structure of the text. For Russian text analysis, we use two approaches: the dependency grammar [3] and the grammar of the direct components [4].

Semantic analysis is related to computer text understanding. There are several popular approaches to semantic analysis:

- semantic role labeling [5] represents semantic roles of words in a sentence through frames;
- entropy based frames as an extension of the role labelling [6];
- compound use of frames with the functional grammar [7];
- semantic text structuring with the generative lexicon [8], [9];

For conversion natural language texts into knowledge representation programming language like Prolog researchers use a semantic analyser preliminary trained on syntactic processing of the natural language texts [10].

The analyser integrates semantic and syntactic analysis in a single procedure or divides on two procedures with results depending on each other. For example: the technology ABBYY Compreno [11] is processing Russian texts in parallel semantically and syntactically affecting each other; in the system ETAP-3 [12], on the first step, the syntactic analyser uses word semantic features and builds the semantic structure on the next step.

Despite of existence of different approaches to the semantic-syntactic text analysis there is a list unresolved problems. The article considers some of them:

- Homonymy resolution [13];
- Synonymy resolution;
- Named entity recognition [14];
- Ellipsis recognition [15];
- Understanding different forms of the same word;
- Metaphor resolution;
- Automatic conversion of the text sense into semantic formalisms for computer processing (understanding);

The article describes an approach to the semantic-syntactic text analysis and automatic filling of the knowledge base, i.e. the implementation of the computer text understanding.

Generally, the process of the natural language text understanding, with the aid of the OSTIS-system, includes next stages [16]:

- Linguistic (graphematic) analysis determines the text components: paragraphs, sentences, words.
- Morphological and syntactic analysis define grammar relations between words in the sentence.
- Semantic analysis generates the sc-text equivalent to the input text.
- Pragmatic analysis integrates the sc-text into the OSTIS-system knowledge base. During the stage:

- - synonymic sc-elements from the knowledge base form pairs with sc-elements of the processed text [17];
  - the terminological system of input text notions aligns with notions of the knowledge base.
- Discourse analysis (context) of the input text. The analysis can have several levels:
  - analysis of neighboring phrases, sentences, and paragraphs (text analysis);
  - context analysis from other sources (extended text analysis). For that, the system should operate with additional information sources: visual and sound. So, a phrase "Look what the man has done" can be analysed and additional information from the visual observation will be integrated;
  - context analysis on the basis of information about source, author, location, and time of the text publication [18].
  - context analysis of the internal knowledge base having some fragment representation of the input text.

Discourse analysis applied on all stages of the text analysis.

The kernel part of the proposed approach is in building the linguistic ontology that integrates: knowledge about linguistics, rules of syntactic and semantic analysis of texts, and the specific subject domain ontology. The ontology includes knowledge about objects and their relations, i.e. gives a formal description of some fragment if a model of the world [22]. Thus, the linguistic ontology includes knowledge about the text and methods of its processing, and the subject domain ontology includes knowledge about some fragment of the real or virtual world, described in the text, and operates with knowledge about rules and methods of specified knowledge processing.

A universal linguistic interface component is built on the bases of the OSTIS technology [19] to be applied in any ostis-system. Every ostis-system consists of the knowledge base in a view of the formal model integrating all knowledge kept in the system and of knowledge processing machine uniting all program agents of the entire system.

SC-code used for formal information representation. Texts of SC-code are interested into the semantic net with basic theory-set interpretation.

The advantages of the technology are following:

- use of unified tools for representing different kinds of knowledge including meta knowledge, which allows to describe all necessary information for analysis;
- use of formalisms allowing to specify as kept notions of the knowledge base so the external computer files;
- provide the ostis-system modifiability, i.e. the ability to extend its functionality.

II. The Structure of the knowledge base of the natural language interface

The knowledge base of the natural language interface has two parts: the language part common for the entire system and the subject part defined by the specific subject domain.

In the ostis-system the subject domain (SD) is a specific structure consisting of:

- main objects of research (OR) – primary and secondary;
- different classes of OR;
- various links where main OR are components, and other types of the links being OR itself having a different level structure;
- different classes of the links (relations);
- different classes of objects not being OR nor the links nor components of the links.

The article considers the subject domain of 'History' represented in the system with sub domains and with corresponded to them ontologies:

- **SD of artefacts** describes all historically valuable and artificially created material entities as a result of a purposeful activity;
- **SD of urban planning** describes immovable monuments of history and culture.
- **SD of persons and social communities** considers a person and all arising from his activity communities of people.
- **SD of ideas** describes compiled on the basis of purposeful activity results [23].
- **SD of historical actions and events** described according to the principles of Semantic coding. [20], [21], [22].

Every SD represented by the set of ontologies [24]:

- **Structure specification** describes roles of the notions and relations of the specific SD with other SDs;
- **Theory-set ontology** describes theory-set relations between the notions of the SD;
- **Logical ontology** includes a system of statements about the notions of the SD;
- **Terminological ontology** represents a system of main and complementary terms (names, signs) corresponding to concepts and relations of the SD and a description of constructing rules of entity terms used as elements (instances) of the concepts and relations.

The constructed knowledge base includes the own knowledge processing machine having program agents implementing logical reasoning based on a hierarchy of statements comprised in the logical ontology.

The linguistic SD represented the language part, which is the common component for all designed systems [25].

Look at a very general structure of the linguistic SD represented in the SCn-language.

SD of Russian language texts
 => *specific SD*:*
 - *SD of Russian language syntax*
 - *SD of Russian language morphology*
 - *Lexical SD of Russian language*

The knowledge-processing machine of the natural language interface
 <= *decomposition of an abstract sc-agent*:*
 {

- *The abstract sc-agent translating external texts into the knowledge base*
- *The abstract sc-agent verifying the knowledge base*

<= decomposition of the abstract sc-agent:*

{

- *The abstract sc-agent verifying correspondence of the relations to its domains*
- *The abstract sc-agent verifying action specification to its class*
- *The abstract sc-agent for searching synonymic elements*

}

- *The abstract sc-agent for context interpretation*

<= decomposition of the abstract sc-agent:*

{

- *The abstract sc-agent identifying the essence correspondent to defined criteria*
- *The abstract sc-agent identifying relations between entities*

}

III. Example of the approach

The example shows the analysis of the following sentences:

- Лангбард И.Г. спроектировал Дом правительства в 1934 году [Langbard I.G. designed the Government house in 1934 year].
- Лангбард И.Г. быд архитектором [Langbard I.G. was an architect].
- Лангбард был архитектором в Комиссии по делам архитектуры [Langbard was an architect in the Commission on Architecture].

Several constraints defined:

- The input of the system are simple narrative Russian sentences;
- The sentence is completed and has a sense;
- Text analysis and understanding implemented for the specific SD (History) and formally represented in knowledge base;
- The knowledge base includes entity signs denoting proper names of persons, establishments, and buildings, which used in the sentences.

Here is a step-by-step analysis of the sentence «Langbard I.G. designed the Government house in 1934 year».

Step 1 Graphematic analysis generates a set of individual words with a given order in the sentence (Fig. 12).

Step 2 The obtained fragments are compared with the samples represented in the knowledge base where they correspond to certain lexemes

Description of lexemes is stored in the linguistic knowledge base. Some of them are on the Figs. 1 - 2

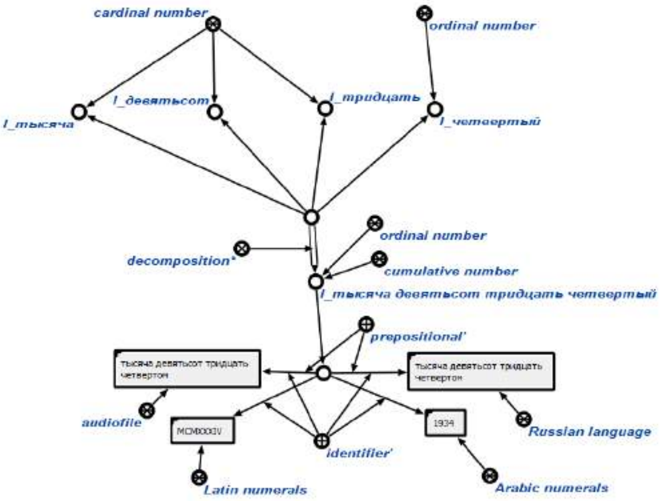

Figure 1. The lexeme '1934'

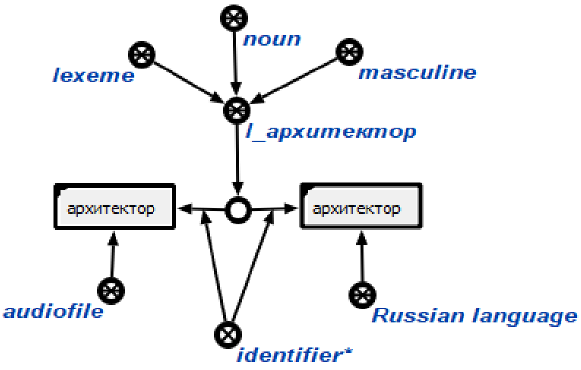

Figure 2. The lexeme 'architect'

The lexemes describing the SD notions at the same time are a part of the linguistic ontology. So, the context of the words described, including it synonyms, through the linguistic ontology.

Named entity recognition is implemented by composing together elements with identical identifiers in the knowledge base. However, naming of the real world objects is much wider of the identification in the knowledge base. Such cases are taken into consideration by the terminological ontology and composition of elements denoting the same object anyway will occur.

Step 3 For each fragment of the text formed a pair «sample-text fragment».

Step 4 The input text is analysed on interrelations between words in the sentence and a syntactic tree is constructed with use of a third-party software [26].

Step 5 The agent translates the result into a semantically equivalent structure in the knowledge base and, as a result, between the words in the sentence appear relations characterising its relationships in the sentence (Fig. 13).

Step 6 In the example, the verb "design" will be found in the historical knowledge base which, in turn, according to the Universal Semantic Code (USC) verb classifier belongs to the class “reproduce” [20], [21].

The ontology of actions is built on the USC basis. The classification of actions supports an idea that different classes of actions differs by the structure of the components.

The USC system satisfies several demands:

- Every USC string corresponds to only one sense;
- Declarative knowledge should be represented in the form of a procedural one. It is important to know not “what is an object in the system", but "what an object performs in the system";
- The means of knowledge representation are not formally separated from the means of knowledge transformation [27].

Thus, each action class has its specific for this action class of performers, mediators, initiators, objects, etc. Based on the characteristics of action classes, the agent is able to transform a verbal description of the action into its formal representation.

Step 7 Homonymy resolution is not applied to the first sample sentence so it will be applied to the two left sentences where accordingly the word ‘architect’ means an occupation and a job position.

In the result of analysis, the agent finds the word ‘architect’ in the historical domain can mean:

1) Subset of the set ‘occupation’

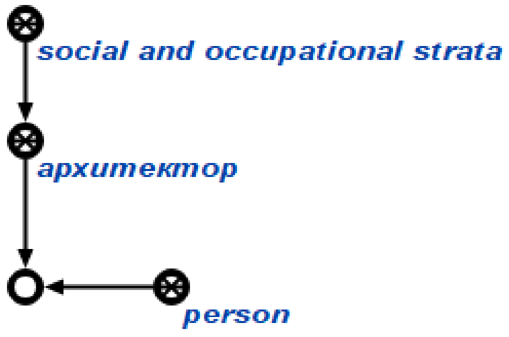

Figure 3. The instance of the class

For the sentence “N was an architect”, without further analysis of words in the sentence, the shown construction will always be formed with the meaning that someone belonging to the class ‘person’ also belongs to the occupation ‘architect’.

2) The role relation ‘architect’, whose domains are the social organisation and the person

The role relation ’architect’ means that some entity of the class ‘person’ has a role of ’architect’ in a scope of the class ‘social organisation’. This construction will be created in further sentence analysis if both relations will be found.

3) Binary relation ’architect’ with domains ‘person’ and ‘building’.

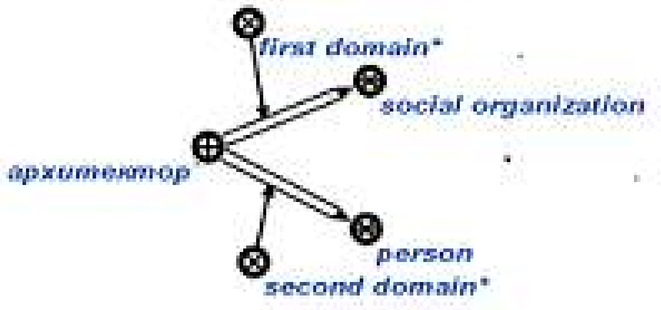

Figure 4. The role relation ’architect’

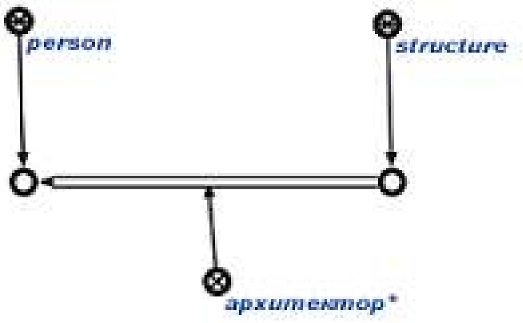

Figure 5. The role relation ’architect’

Working with first sentence the agent is able to find Langbard I.G. belongs to the set ‘architect’ and to form the correspondent structure in the knowledge base.

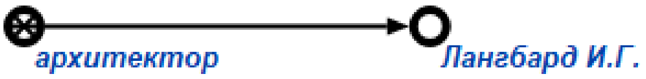

Figure 6. Representation of the sentence “Langbard I.G. was an architect”

For the second sentence the construction on the Fig.6 will be formed again and because the construction already exists in the knowledge base the synonymic elements will be composed together based on the identity of its identifiers and roles in the formalism.

Then, in the sentence will be found notions belonging to the class ‘person’ – Langbard I.G. and to the class ‘social organisation’– Commission on Architecture. Therefore, the next construction will be formed Fig. 7.

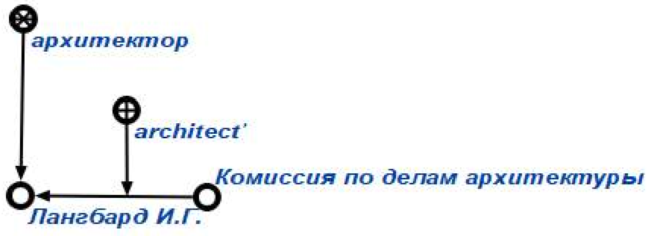

Figure 7. Representation of the sentence “Langbard was an architect in the Commission on Architecture”

The construction will be added to the knowledge base as a unique element and became a part of the entity description named Langbard I.G.

For the binary relation ‘architect’ the second data domain belonging to the class ‘building’ will not be found and the agent ends functioning because there are no other variants of use of the notion ‘architect’.

Thus, the homonymy resolution in the scope of the system implemented when specifying notions. *Step 8* In a result of transforming the text into an equivalent formal structure, it becomes possible to derive logical results based on the data available in the sentence. It is the essence of the context analysis.

The further work of the agent for revealing the relations between entities uses the hierarchy of statements described in the history SD, while the search of statements will be limited only with concepts extracted within the semantic-syntactic analysis of the text, or resulting from contextual analysis.

Accordingly, the set of logical rules will be finite and limited with the Logical ontology of the specific SD. The effectiveness and completeness of the contextual analysis depends directly on the completeness of ontologies.

Consider the results of text contextual analysis for the proposed fragment ‘Government House’ within the knowledge base.

Initially, the phrase "Government House" has already been correlated in the knowledge base on history with the set ‘building’. The same will happen with the term ‘government’. As a result, the following structure is formed.

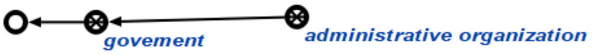

Figure 8. Determining the notions through the knowledge base on history

The construction means: there is some entity that belongs to the set ‘government’, which in turn is a subset of the set ‘administrative organisation’.

Then, the agent analyses and selects all types of relations existing between instances of the sets ‘building’ and ‘organisation’, and alternately applies existing statements for the found relations.

Fig. 9 demonstrates the rule for analysis possible relations between the notions ‘government’ and ‘building’.

After applying this rule, the existing structure in the knowledge base is supplemented with new information.

Fig. 10 demonstrates as the existing structure was supplemented with information that with a probability of 0.3 the government owned the building called "Government House", and with the probability 0.7 the government was located in the building.

The coefficient of probability is calculated by counting the frequency of used relations between the entities, where "1" is the total number of all found relations, and the coefficient is the fraction of each relation in the total number. The coefficient reflects the current state of the knowledge base on history, and can change with its further completion.

Further, the proposed changes should be approved and placed into the knowledge base or rejected.

Step 9 Then, the agent for identifying the entity that meets the specified criteria is activating.

Having found the date in the sentence, the agent is replacing the node without the identifier belonging to the set ‘government’, with the node with the identifier "Council of People’s Commissars of the BSSR", composing it together with the corresponding elements in the knowledge base on the ground that in 1934 the Council of People’s Commissars of the BSSR played the role of the government (Fig11)

As a result of the analysis, all formed structures will become the part of the knowledge base, and the historical system will have enough knowledge to answer following questions:

- Who designed the Government House?
- What is the Government House?
- What is the government?
- What was located in the Government House?
- Who was the owner of the Government House?
- Which organisation acted as the government in 1934?
- Who was Langbard I.G.?

IV. Conclusion

The proposed approach of the text analysis has the following advantages:

- All stages of the text analysis implemented in the same environment, what excludes the need to address compatibility issues between different solutions;
- Semantic net fragments equivalent to the text become the result of the text analysis, and text become the part of the knowledge base and this allows further processing of the received knowledge by standard methods, and also allows to organise the automatic filling of the knowledge base from texts;
- Common use of linguistic knowledge bases and any other SD allows to resolve problems of homonymy and synonymy easier and faster;
- The availability of knowledge bases for the specific SD allows use of logical ontologies for contextual analysis of the text within the given SD.

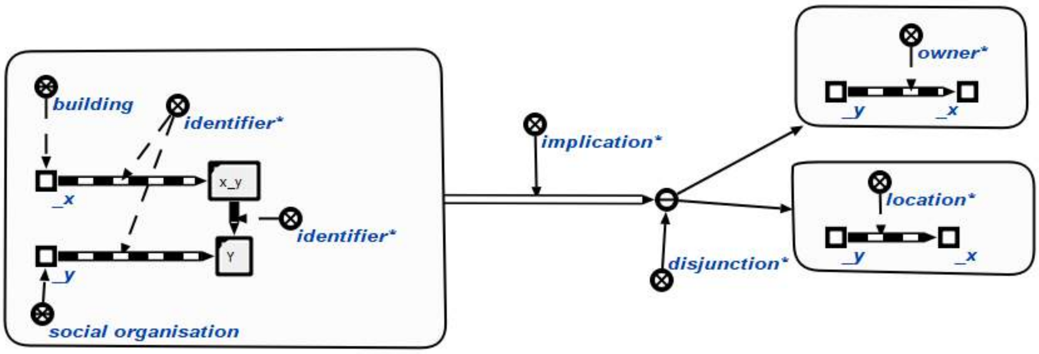

Figure 9. The logical rule found in the knowledge base

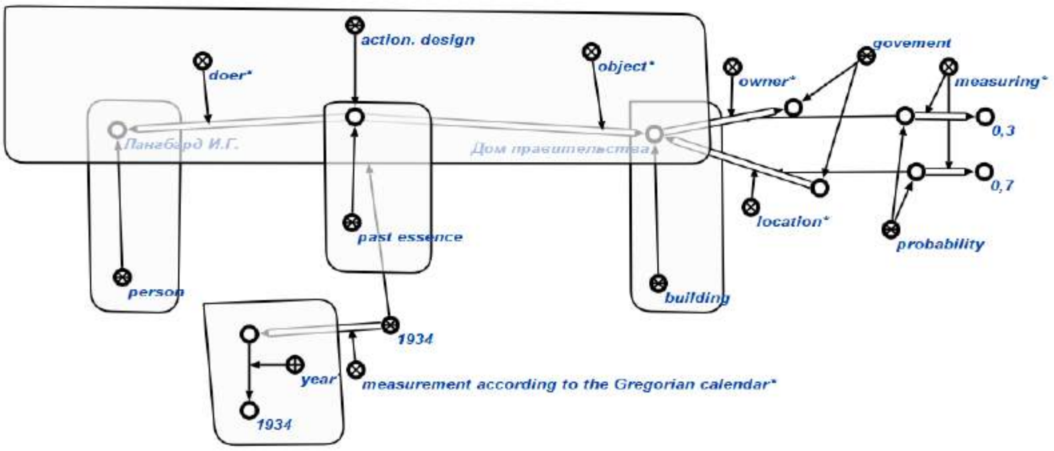

Figure 10. Functioning of the context analysis agent

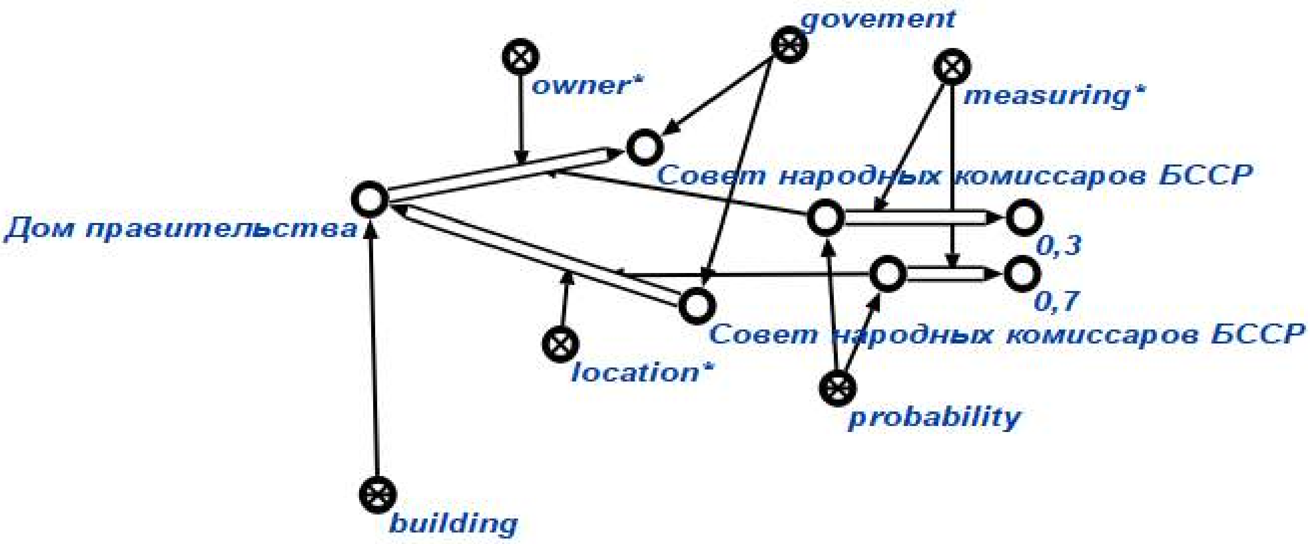

Figure 11. Synonymic elements after the agent functioning

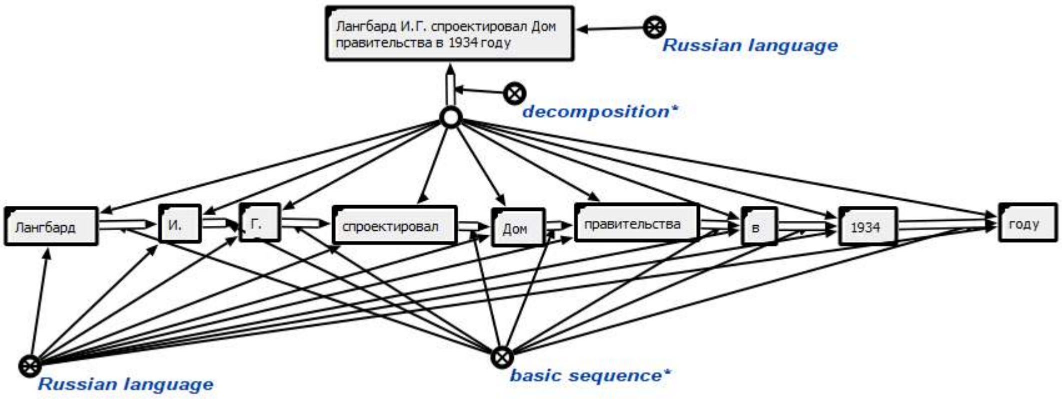

Figure 12. Graphematic analysis of the input sentence

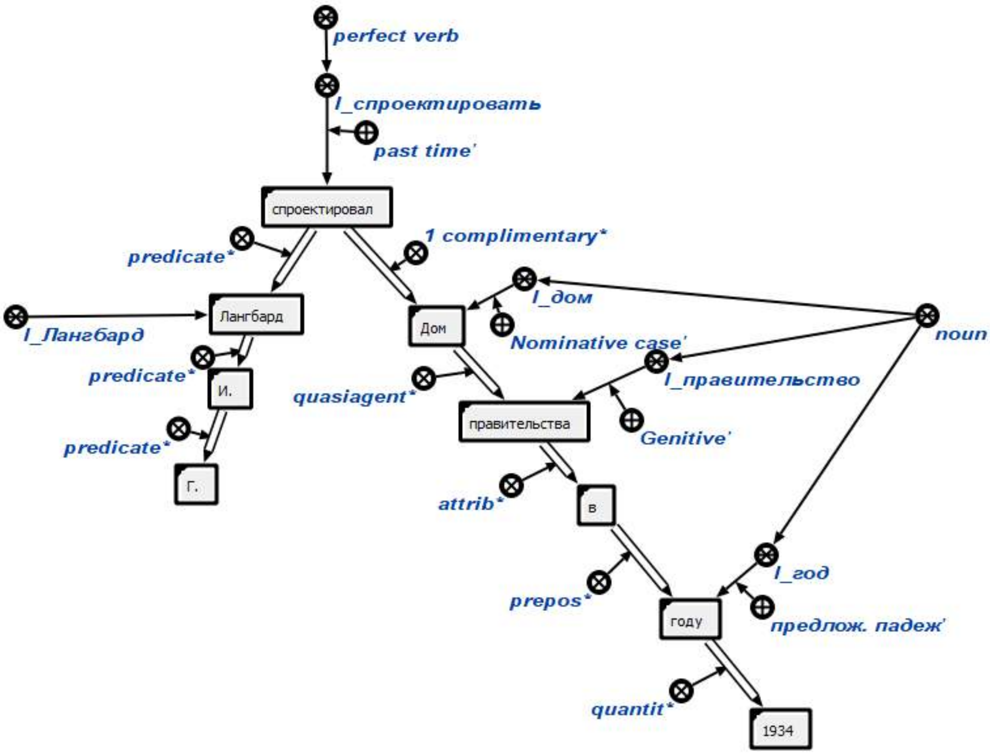

Figure 13. The result of translating the tree constructed by the ETAM service into a semantically equivalent structure in the knowledge base

References

[1] Tarasov V.B. The Problem of Understanding: The Present and Future of Artificial Intelligence. Open semantic technologies for intelligent systems, 2015, pp. 25-42.

[2] Pospelov D.A. Ten hot spots in research on artificial intelligence./ Intellectual systems, 1996, vol. 1,No 1-4, pp. 47-56.

[3] Zolotova GA, Onipenko NK, Sidorova M. Yu. Communicative grammar of the Russian language. Moscow, Nauka, 2004. 544 p/

[4] Gladkiy A.V. Formal grammars and languages. Moscow, Nauka, 1973. 386 p.

[5] Gildea D., Jurafsky D. Automatic labelling of semantic roles. Comput. Linguist, 2002, Vol. 28, no. 3, pp. 245–288.

[6] Bharati A., Venkatapathy S., Reddy P. Inferring semantic roles using sub-categorisation frames and maximum entropy model. Proceedings of the Ninth Conference on Computational Natural Language Learning, 2005, pp. 165–168.

[7] De Busser R., Moens M.-F. Learning Generic Semantic Roles. Available at: http://www.academia.edu/265713/Learning_Generic_Semantic_Roles

[8] Claveau V., Sebillot P. From efficiency to portability: acquisition of semantic relations by semi-supervised machine learning. Proceedings of COLING'04, 20th International Conference on Computational Linguistics, 2004, pp. 61–267.

[9] Ichiro Yamada T. B. Automatic discovery of telic and agentive roles from corpus data. Proceeding of the 18th Pacific Asia Conference on Language, Information and Computation, 2004, pp. 115-126.

[10] Mooney R. J. Learning for semantic parsing. Computational Linguistics and Intelligent Text Processing: Proceedings of the 8th International Conference, 2007, pp. 311– 324.

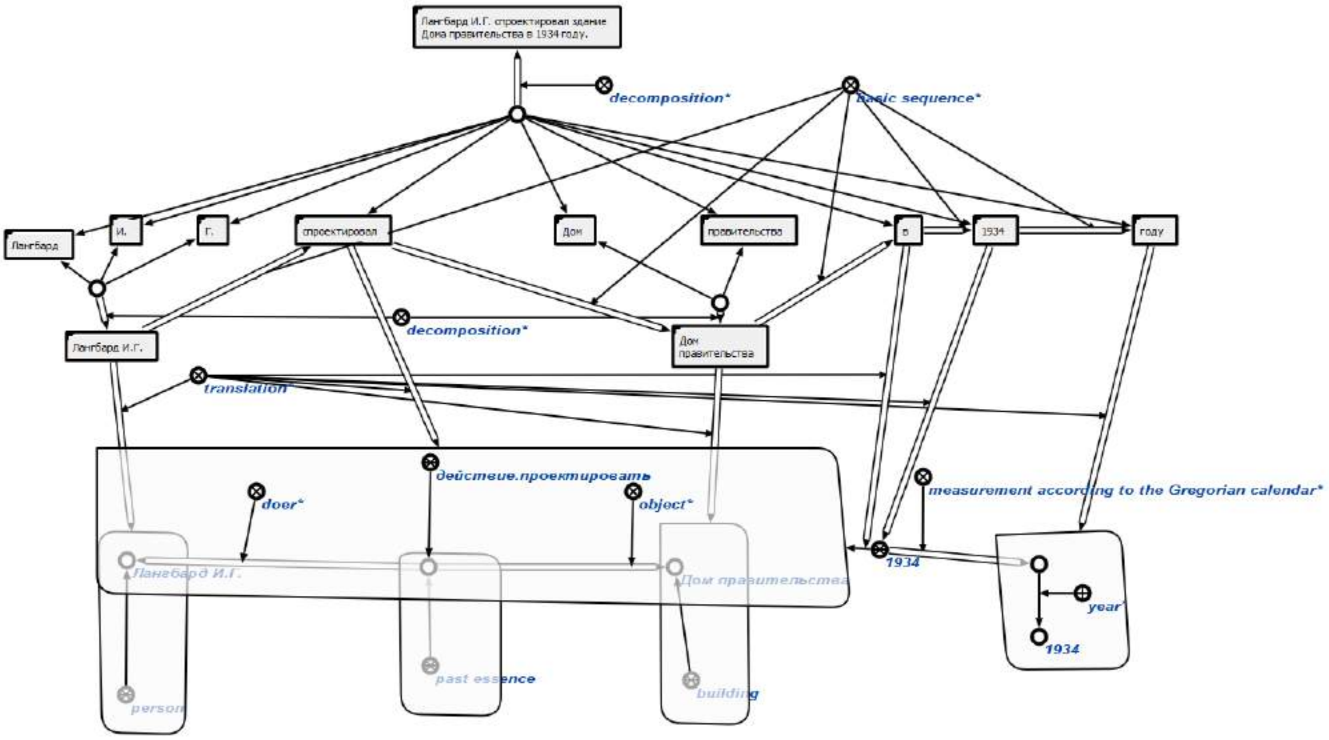

Figure 14. The result of translating the natural language text into a structure in the language of the sc-code

[11] Anisimovich K. V.,Druzhkin K. Ju Syntactic and semantic parser based on ABBYY Compreno linguistic technologies. Papers from the Annual International Conference "Dialogue" , 2012, vol. 2., pp. 91–103.

[12] Iomdin L., Petrochenkov V. ETAP parser: state of the art. Papers from the Annual International Conference "Dialogue", 2012, pp. 830–853.

[13] Anjali, M K., Babu, Anto P. 2014. Ambiguities in Natural Language. Processing International Journal of Innovative Research in Computer and Communication Engineering, 2014, Vol.2, Special Issue 5, pp. 392–394.

[14] Mozharova V.A., Lukashevich N.V. Research of signs for extracting named entities from texts in Russian. Scientific and technical information. Series 2: Information Processes and Systems, 2017, No. 5, pp. 14-21.

[15] Tonoyan-Belyaev I.A. Ellipse and the concept of conventional syntax. Bulletin of St. Petersburg University, 2007, Series 6, Issue. 3, pp. 232-236.

[16] Maksimov V.Yu., Klyshinsky ES The problem of understanding in artificial intelligence systems. New information technologies in automated systems, 2016, pp. 43-60

[17] Ivashenko V.P. Models and algorithms of knowledge integration based on homogeneous semantic network. Dokt, Diss., Minsk, 2014. 147 P.

[18] Pospelov D.A. Levels of understanding. Artificial intelligent. Book 2. Models and methods: Manual, Moskow, Radio and Communication, 1990, pp.110-115

[19] The IMS.OSTIS website. Available at: http://www.ims.ostis.net.

[20] Martynov V.V. Fundamentals of semantic coding. Experience in the representation and transformation of knowledge.Minsk, 2001. 138 p.

[21] Boyko.I.M. Semantic classification of actions for knowledge inference. Open semantic technologies for intelligent systems, 2016, pp. 115-121

[22] Hardzei A.N. Theory of the architecture automatic formation (TKAAF-2) and futher minimisation of semantic calculus. . Open semantic technologies for intelligent systems, 2014, pp.49-64

[23] Hubarevich A.V. Ontology-Based Design of Intelligent Systems in the Field of History. Open semantic technologies for intelligent systems, 2017, pp. 245-250.

[24] Davydenko I.T., Grakova N.V. Means of structuring of semantic models kazan knowledge. Open semantic technologies for intelligent systems, 2016, pp.93-107

[25] Rusetskiy K.V. Approaches to the natural language formatting in the semantic graph language of knowledge representation. Karpov Scientific Readings: a collection of scientific articles, iss. 7, part 1., pp. 255-259.

[26] Laboratory of Computational Linguistics. Available at: http://cl.iitp.ru/ru/etap3

[27] Martynov V.V. Semantic coding for representation and knowledge transformation.Moscow, Institute for Information Transmission Problems of the USSR Academy of Sciences, 16p.

ПРИОБРЕТЕНИЕ ЗНАНИЙ НА ОСНОВЕ ТЕКСТОВ ЕСТЕСТВЕННОГО ЯЗЫКА

Бойко И.М., Гордей А. Н.
Центр системного анализа и стратегических исследований НАНБ
г. Минск, Беларусь
Губаревич А.В., Семеняка А.Ф.
Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Беларусь

В статье описывается подход к машинному анализу естественно-языкового текста с последующим автоматическим наполнением базы знаний на основе технологии OSTIS. Данный подход позволяет проводить семантико-синтаксический анализ текстов с последующим анализом контекста, что достигается за счет построения онтологий конкретной предметной области.